On Effective Learning for Multimodal Data

by

Tanzila Rahman

M.Sc., University of Manitoba, 2018

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Computer Science)

The University Of British Columbia

(Vancouver)

April 2024

© Tanzila Rahman 2024
The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

On Effective Learning for Multimodal Data

Submitted by ___________Tanzila Rahman__________ in partial fulfillment of the requirements for the degree of ___________Doctor of Philosophy__________

in ___________Computer Science__________

Examining Committee:

Leonid Sigal, Professor, Computer Science, UBC
Supervisor

James J. Little, Professor, Computer Science, UBC
Supervisory Committee Member

Kwang Moo Yi, Assistant Professor, Computer Science, UBC
Supervisory Committee Member

Michiel Van De Panne, Professor, Computer Science, UBC
University Examiner

Zhen Jane Wang, Professor, Electrical and Computer Engineering, UBC
University Examiner

Vicente Ordóñez Román, Associate Professor, Computer Science, Rice University
External Examiner
Abstract

Humans can perceive the world through multiple modalities. Strong behavioral scientific evidence suggests that such ability, which includes implicit information integration and cross-modal alignment inherent in it, is critical for human learning. Nevertheless, until relatively recently, most deep learning methods have primarily focused on addressing single-modality issues associated with learning from vision, sound, or text. Over the recent years, however, researchers started to focus on multi-modal learning, specifically emphasizing high-level visual comprehension challenges like image-text matching, video captioning, and generation of audio-visual content. In this thesis, we aim to broaden the scope of learning from multi-modal information, enhance its integration, and solve problems related to human-centric spatio-temporal perception in a manner that does not necessarily require complete supervision (e.g., granular spatio-temporal multi-modal alignment).

Specifically, we focus on addressing two fundamental challenges: (1) Multi-modal learning; and (2) Weak-supervision. We address these challenges across a range of diverse tasks. First, we focus on weakly-supervised dense video captioning, where we combine audio with visual features to improve state-of-the-art performance. We also show that audio itself can carry a surprising amount of information, compared to existing visual-only models. Secondly, we introduce an end-to-end audio-visual co-segmentation network to recognize individual objects and corresponding sounds using only object labels, without requiring any additional supervision or bounding box proposals. Third, we propose TriBERT, a transformer-based architecture with co-attention, that learns contextual features across three modalities: vision, pose, and audio. We show that these features are general and improve performance on a variety of tasks spanning audio-visual sound source separation and cross-modal retrieval. Fourth, we delve into generative text-to-image (TTI) models, specifically to address consistency when generating complex story visualizations by augmenting diffusion models with memory module. Finally, we look at aspects of penalization within TTI. This allows us to generate diverse visuals for custom and user-specified concepts (e.g., a specific person, dog, etc.).

Throughout our comprehensive analysis of these tasks within this thesis, we present significant algorithmic, theoretical, and empirical contributions to the field of multimodal machine learning and computer vision.
Lay Summary

This thesis explores enhancing artificial intelligence for autonomous systems by integrating various senses like vision, language, and audio, inspired by how humans learn from their environment. Unlike traditional approaches that mostly address single-sensory information, this research focuses on combining and integrating information from multiple modalities. The primary goal of this thesis is to enhance learning from diverse sources and tackle challenges related to human-centric perception over time and space. To this end, we identify two key challenges to build multimodal autonomous systems – multi-modal learning and weak supervision. We address these challenges and develop innovative approaches such as weakly-supervised dense captioning, audio-visual co-segmentation and localization, and a transformer-based architecture called TriBERT. The research also contributes to generative text-to-image models for creating consistent story visualizations and generating diverse visuals based on user-specified concepts. Overall, the thesis substantially advances multimodal machine learning and computer vision with novel and practical contributions.
Preface

This thesis is submitted in partial fulfillment of the requirements for a Doctor of Philosophy degree in Computer Science. The whole research presented here was conducted by me, Tanzila Rahman, along with other co-authors, under the supervision of Professor Leonid Sigal.

Chapter 3. A version of this chapter has been published by Rahman, T., Xu, B., Sigal, L., “Watch, listen and tell: Multi-modal weakly supervised dense event captioning”, Proceedings of the IEEE/CVF international conference on computer vision, 2019 [258]. I came up with the idea, designed and implemented experiments, and conducted all the experiments. I also wrote and presented the paper. Bicheng Xu assisted with writing the paper and contributed to discussions and analysis.

Chapter 4. A version of this chapter has been published by Rahman, T., Sigal, L., “Weakly-supervised audio-visual sound source detection and separation”, IEEE International Conference on Multimedia and Expo, 2021 [256]. I was the sole student author responsible for developing concepts, implementing them, and writing the paper.

Chapter 5. This chapter was published by Rahman, T., Yang, M., Sigal, L., “TriBERT: Human-centric audio-visual representation learning”, Advances in Neural Information Processing Systems, 2021 [304]. I was the main author, handling concept formation, experiments, analysis, and most writing. Mengyu Yang contributed to the implementation and writing of the multi-modal retrieval section.

Chapter 6. This chapter was published by Rahman, T., Lee, H. Y., Ren, J., Tulyakov, S., Mahajan, S., Sigal, L., “Make-a-story: Visual memory conditioned consistent story generation”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023 [257]. I developed the idea, researched prior work, and designed the structure and framework. Shweta Mahajan assisted with the analysis and contributed to writing the paper. Lee, H. Y., Ren, J., and Tulyakov, S. contributed to the project by participating in discussions and analyzing the results.

Chapter 7. This chapter is currently under review for a conference. I conceived the concept, reviewed previous studies, developed and implemented the architecture, and composed and presented the paper. Shweta Mahajan contributed to the analysis, review of previous studies, and assisted in writing the paper.
# Table of Contents

Abstract .................................................. iii

Lay Summary .............................................. iv

Preface .................................................... v

Table of Contents ........................................ vi

List of Tables ............................................ ix

List of Figures ........................................... xi

List of Abbreviations .................................... xvii

Acknowledgements ....................................... xix

Dedication ............................................... xx

## 1 Introduction ....................................... 1

1.1 Motivation and Objectives .......................... 2

1.2 Thesis Outline and Contributions ............... 4

1.3 Publications ........................................ 8

## 2 Background ........................................... 10

2.1 Foundations of Multimodal Learning ............ 10

2.2 Core Principles in Multimodal Research ........ 14

2.2.1 Modalities are Diverse ....................... 14

2.2.2 Modalities are Interconnected ............... 15

2.2.3 Modalities Interact ......................... 15

2.3 Applications of Multimodal Learning: ........ 16
<table>
<thead>
<tr>
<th>Table of Contents</th>
<th>vii</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Multi-modal Weakly Supervised Dense Event Captioning</td>
<td>18</td>
</tr>
<tr>
<td>3.1 Related Work</td>
<td>20</td>
</tr>
<tr>
<td>3.2 Multi-modal Dense Event Captioning</td>
<td>21</td>
</tr>
<tr>
<td>3.2.1 Feature Representation</td>
<td>23</td>
</tr>
<tr>
<td>3.2.2 Weakly Supervised Model</td>
<td>24</td>
</tr>
<tr>
<td>3.2.3 Context Fusion</td>
<td>26</td>
</tr>
<tr>
<td>3.2.4 Training Loss</td>
<td>28</td>
</tr>
<tr>
<td>3.3 Experiments</td>
<td>28</td>
</tr>
<tr>
<td>3.3.1 Dataset</td>
<td>28</td>
</tr>
<tr>
<td>3.3.2 Experiment Setup and Implementation Details</td>
<td>28</td>
</tr>
<tr>
<td>3.3.3 Experiment Results</td>
<td>31</td>
</tr>
<tr>
<td>3.4 Conclusion</td>
<td>33</td>
</tr>
<tr>
<td>4 Weakly-supervised Audio-visual Sound Source Detection and Separation</td>
<td>34</td>
</tr>
<tr>
<td>4.1 Related Work</td>
<td>36</td>
</tr>
<tr>
<td>4.2 Approach</td>
<td>37</td>
</tr>
<tr>
<td>4.2.1 Problem Formulation</td>
<td>37</td>
</tr>
<tr>
<td>4.2.2 Weakly-supervised Audio-Visual Architecture</td>
<td>38</td>
</tr>
<tr>
<td>4.3 Experiments</td>
<td>41</td>
</tr>
<tr>
<td>4.3.1 Datasets</td>
<td>41</td>
</tr>
<tr>
<td>4.3.2 Pre-processing and implementation details</td>
<td>42</td>
</tr>
<tr>
<td>4.3.3 Experimental results for source separation and detection</td>
<td>43</td>
</tr>
<tr>
<td>4.4 Conclusion</td>
<td>47</td>
</tr>
<tr>
<td>5 Human-centric Audio-visual Representation Learning</td>
<td>48</td>
</tr>
<tr>
<td>5.1 Related works</td>
<td>50</td>
</tr>
<tr>
<td>5.2 Approach</td>
<td>51</td>
</tr>
<tr>
<td>5.2.1 Reviewing Vision-and-Language BERT (ViLBERT)</td>
<td>51</td>
</tr>
<tr>
<td>5.2.2 TriBERT Architecture</td>
<td>53</td>
</tr>
<tr>
<td>5.3 Experiments</td>
<td>59</td>
</tr>
<tr>
<td>5.3.1 Experiments for Sound Separation</td>
<td>59</td>
</tr>
<tr>
<td>5.3.2 Multi-modal Retrieval</td>
<td>60</td>
</tr>
<tr>
<td>5.4 Conclusion</td>
<td>65</td>
</tr>
<tr>
<td>6 Visual Memory Conditioned Consistent Story Generation</td>
<td>66</td>
</tr>
<tr>
<td>6.1 Related work</td>
<td>68</td>
</tr>
<tr>
<td>6.2 Approach</td>
<td>71</td>
</tr>
<tr>
<td>6.2.1 The Latent Diffusion Model Backbone</td>
<td>71</td>
</tr>
<tr>
<td>6.2.2 Story-Latent Diffusion Models</td>
<td>72</td>
</tr>
</tbody>
</table>
# Table of Contents

6.3 Datasets and Evaluation Metrics ........................................ 77  
6.4 Experiments ................................................................. 79  
6.5 Additional Quantitative Results .......................................... 83  
6.6 Additional Qualitative Results ........................................... 85  
6.7 Conclusion ........................................................................ 85  

7 Visual Concept Driven Image Generation with Text-to-Image Diffusion Model ......................................................... 89  
7.1 Introduction ....................................................................... 89  
7.2 Related Work .................................................................... 92  
7.3 Methodology ..................................................................... 94  
7.4 Experimental Results ........................................................ 99  
7.5 Limitations and Future Directions ....................................... 109  
7.6 Conclusion ........................................................................ 110  

8 Conclusions ........................................................................ 111  
8.1 Research summary and impact ........................................... 111  
8.2 Extensions and perspectives of my research ....................... 113  
8.3 Ethics Assessment ............................................................. 114  

Bibliography .......................................................................... 116
## List of Tables

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td><strong>Audio Only Results.</strong></td>
<td>Illustrated are dense captioning results of pre-trained and final models using audio only.</td>
<td>29</td>
</tr>
<tr>
<td>3.2</td>
<td><strong>Fusion Strategies.</strong></td>
<td>Testing results for different context fusion strategies for integrating audio and video modalities are illustrated for both pre-trained and final models. We use MFCC audio features and C3D video features for all experiments.</td>
<td>29</td>
</tr>
<tr>
<td>3.3</td>
<td><strong>Multi-modal Results.</strong></td>
<td>Comparison among unimodal and our multi-modal models using MUTAN fusion.</td>
<td>30</td>
</tr>
<tr>
<td>3.4</td>
<td><strong>Results with ground-truth temporal segments.</strong></td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>3.5</td>
<td><strong>Pretrained model results on the full dataset.</strong></td>
<td></td>
<td>31</td>
</tr>
<tr>
<td>4.1</td>
<td><strong>Audio separation results on MUSIC test set.</strong></td>
<td>Performance reported using SDR/SIR/SAR. SDR and SIR capture separation accuracy; SAR only captures the absence of artifacts.</td>
<td>43</td>
</tr>
<tr>
<td>4.2</td>
<td><strong>Multi-label object classification accuracy.</strong></td>
<td>Performance in (%) on the MUSIC test set.</td>
<td>44</td>
</tr>
<tr>
<td>4.3</td>
<td><strong>Cross dataset evaluation of audio separation.</strong></td>
<td>Evaluation of the model trained using the MUSIC dataset on the AudioSet-SingleSource dataset; using ratio mask. SAR is responsible for capturing absence of artifacts, therefore, if can be higher even when separation results are poor.</td>
<td>46</td>
</tr>
<tr>
<td>4.4</td>
<td><strong>Audio separation for unseen objects.</strong></td>
<td>Toy experiment with cross-dataset setting where the model has never seen some instruments during training on MUSIC dataset.</td>
<td>46</td>
</tr>
<tr>
<td>5.1</td>
<td><strong>Sound separation results on the MUSIC test set.</strong></td>
<td>We use TriBERT with pre-trained weights from the MUSIC21 dataset and then fine-tune this model using the MUSIC train set.</td>
<td>60</td>
</tr>
<tr>
<td>5.2</td>
<td><strong>Sound separation results on the MUSIC21 test set.</strong></td>
<td>SDR / SIR / SAR are used to report performance. Separation accuracy is captured by SDR and SIR; SAR only captures the absence of artifacts.</td>
<td>60</td>
</tr>
</tbody>
</table>
5.3 Multi-modal retrieval on the MUSIC21 test set. Top-\(k\) accuracy results for 5 retrieval variants. For each variant, retrieval on both the transformer-based (bolded) and pre-transformer (un-bolded) embeddings were evaluated. Also shown is the accuracy of a random selection model. ........................................... 64

6.1 Dataset statistics of the MUGEN, FlintstonesSV and PororoSV. 73
6.2 Quantitative results. Experimental results on the FlintstoneSV and the MUGEN datasets. ................................. 73
6.3 Experimental results using our proposed Story-LDM with and without the reference text on the FlintstonesSV dataset. 83
6.4 Quantitative results. Experimental results on the PororoSV datasets. 83
6.5 Experimental results using non-referential text on the Flintstones and mugen dataset. ................................. 85

7.1 Quantitative evaluation in terms of image similarity and diversity on COCO dataset. ................................. 100
7.2 Quantitative evaluation based on user study. ...................... 101
List of Figures

3.1 Multi-modal Dense Event Captioning. Illustration of our problem definition, where we use both audio features and visual information to generate the dense captions for a video in a weakly supervised manner. .......................................................... 19

3.2 Our Multi-modal Architecture. The model has two parts, a sentence localizer and a caption generator. The sentence localizer takes audio, video, and captions as inputs and generates a temporal segment for each caption. The caption generator uses the resultant temporal segments, with audio and video features, to produce a caption for each segment. .......................................................... 22

3.3 Context Fusion Strategies. Three fusion strategies are illustrated: (a) multiplicative mixture fusion, (b) multi-modal context fusion, and (c) MUTAN fusion.......................................................... 24

3.4 Qualitative Results. Both pretrained and final model results are illustrated of two videos. Captions are from (a) ground-truth; (b) pretrained model trained only using visual features; (c) multi-modal pre-trained model; (d) the final model trained with video features only; (e) our multi-modal final model for dense event captioning in videos.......................................................... 32

4.1 Visually-guided Sound Source Separation. Our method first detects and segments objects which make sounds, in a weakly-supervised manner, and then separates their respective audio signals. ....... 38
4.2 **Weakly-supervised Audio-Visual Architecture.** ResNet-18 followed by a $3 \times 3$ convolution layer is used to extract visual feature ($V_f$) from input video frames. These features are fed to the segmentation network to detect the sound sources. Depending on the classification scores from the segmentation network, we generate soft semantic segmentations by producing a class-specific attention map ($X_m$). We use this attention map to pool features from respective image regions of $V_f$ generating $V_{fm}$. The resultant feature is concatenated with the bottleneck features of attention u-net to generate an audio-visual feature ($W_{AV}$). $W_{AV}$ is passed to the mask coefficient generator to generate $k$ mask coefficients ($M$). At the same time, attention U-Net generates $k$ audio channels ($P$) and combined linearly ($\sigma(PM^T)$) to predict the final audio spectrogram mask guided by visual feature.

4.3 **Qualitative audio separation results on MUSIC test set.** Test samples, our results and comparison with [84] are shown. See text for details and discussion.

4.4 **Attended object map.** Attended map with class labels from our learned weakly supervised segmentation network which focus on objects that make sounds.

5.1 **Our TriBERT Architecture.** We train TriBERT on the MUSIC21 dataset under two training tasks: (1) Classification and (2) Sound separation. We introduce an end-to-end segmentation network for visual embeddings which takes consecutive RGB frames as input and outputs detected object features to feed into vision BERT. Following [79], we use graph CNN to generate pose embeddings as input to pose BERT. For audio, we consider the mixed spectrogram of two audio sources and use a VGGish network to generate audio embeddings to feed into audio BERT. We use classification loss to train individual modalities. Due to the lack of annotation for individual objects, we use a weakly supervised classification loss to train TriBERT for the vision and pose streams. Following prior works on sound separation, we utilize attention U-net [227], which takes a mixed audio spectrogram as input and predicts a spectrogram mask guided by audio-visual-pose features.
5.2 **Visual Segmentation Network.** We consider ResNet50 [107] followed by a $3 \times 3$ convolution as our backbone network. It outputs a $H \times W$ spatial visual feature ($V_e$) fed into the segmentation network. Class specific attention map ($A_m$) is generated by the segmentation network and used to pool top two detected object features. We consider the features as visual embeddings for TriBERT. 54

5.3 **Tri-modal Co-attention Layer.** The key and value matrices from two modalities are concatenated together and feed as an input to other modality. The final output of Tri-modal co-attention layer is multi-modal feature for each modality conditioned on other two modalities. 56

5.4 **Qualitative sound separation results on the MUSIC21 test set.** Here, we show a comparison between our results and Music-Gesture [79]. See text for details. 61

5.5 **Qualitative sound separation results on the MUSIC21 test set.** Here, we show a comparison between our results and Music-Gesture [79]. 62

5.6 **Overview of the training scheme for the cross-modal retrieval module in the $n = 4$ case.** Each multiple choice consists of the correct vision+audio fusion embedding along with a pose embedding. The 4 choices: 1 containing the correct pair, 2 containing easy negatives, and 1 containing a hard negative is passed as input into a MLP network. The network outputs a score for each pair, after which a softmax is applied across all 4 scores. We train this network in a siamese fashion (i.e. each MLP block in the figure shares the same weights) using cross entropy loss. In our experiments, we use a 4-layer MLP with Tanh activation. 63

5.7 **Qualitative cross-modal retrieval results on the MUSIC21 test set.** The similarities (same instrument class) between results in the top-5 retrieval pool show that our transformer-based representations have learned a semantically meaningful relationship between the modalities. 65

6.1 **Referential and consistent story visualization.** Examples of more natural stories with references for the FlintstonesSV [97] and MUGEN [104] datasets (bottom text) compared to more typical but less natural story text (top). We extend the MUGEN dataset by introducing additional two characters (*e.g.* Lisa and Jhon) and four backgrounds (*e.g.* Sand, Grass, Stone, and Dirt). 66
6.2 **Story-Latent Diffusion Models for consistent story generation.**  
*(a)* The autoregressive conditional generative Story-LDM model with the proposed memory-attention module. *(b)* A snapshot of the memory for $k$ frames. ........................................ 69

6.3 **Story generation result for MUGEN.** Here we can compare our method and LDM [265]. See text for details. ................. 74

6.4 **Story generation results for FlintstoneSV.** Our method is able to generate more consistent characters/backgrounds. ............... 75

6.5 **Qualitative Comparison on Story Generation.** Comparison on the FlintstonesSV dataset visual story generation. .......... 78

6.6 **Comparison to Story Continuation.** Comparison on the FlintstonesSV dataset for story continuation. Prior work uses the first frame as additional input to the model; our model does not. ....... 79

6.7 **Story Diversity.** Diverse outputs for a single storyline obtained with our Story-LDM. ................................. 80

6.8 **Branching Storyline.** Generating different yet consistent stories by branching the storyline. Frames in the later columns are generated based on earlier ones and corresponding text. .............. 82

6.9 **Combination of different characters and backgrounds from the MUGEN dataset [104].** ................................. 83

6.10 **Positive (1st row) and negative (2nd row) examples.** .............. 84

6.11 **Qualitative results on the MUGEN dataset [104].** ............... 86

6.12 **Qualitative results on the FlintstonesSV dataset [97].** ........... 87

6.13 **Qualitative results on the PororoSV dataset [186].** ............... 88

7.1 **Concept-driven image generation:** Given images depicting multiple concepts (*subjects and context/background*), the top output is the illustration of the male ([v1]) being generated in the context/background of the female ([v3]) image by different methods. The bottom output illustrates the two concepts together in a single image. Dreambooth [270] (left) encodes [v1], [v2], [v3] from multiple input concept images. It fails to generate multi-concept interactions. Break-a-scene [16] (middle) disentangles [v1], [v2], [v3] from single image. This approach requires human-annotated masks. Our approach (right) disentangles [v1], [v2], [v3] from single image. The *latent masks* are obtained from EM-like optimization. Using these optimized masks, our method can automatically produce images with those concepts in new contexts either by themselves or, jointly, interacting with one another. ................................. 90
7.2 **Overview of our EM-style Optimization Framework.** Unique tokens are first assigned to the concepts. In this example, three tokens are assigned: [v1] to the male subject in the first image, [v2] to the female subject in the second image, and [v3] to the background in the second image. Where appropriate these tokens could be initialized with CLIP text embeddings semi-representative of the concept (e.g., “person” for both [v1] and [v2]) or simply initialized with random vector embeddings. The latent masks are then initialized by averaging cross-attention maps (see left of Step 1) between the newly defined tokens and the corresponding images across randomly selected 50 diffusion timesteps. The resulting attention maps are subsequently binarized and refined using Dense-CRF, resulting in latent binary masks (see right of Step 1). In Step 2, the tokens are re-optimized for a newly sampled timestep with a combination of Mask Diffusion Loss (3) and Cross-Attention Loss (4) [see text for details]. The new tokens are then used to refine cross-attention and the mask. This alternating optimization continues for a fixed number of steps until masks and tokens converge.

7.3 **Quantitative comparison - Mask IoU as a function of training steps.** We compare the estimated (latent) mask and the manually annotated mask for the concepts illustrated in Figure 7.1. Performance, as a function of mask optimization, for different tokens over training steps is illustrated: (a) Mask optimization for Token1 (i.e. subject), (b) Mask optimization for Token2 (i.e. another subject), and (c) Mask optimization for background Token. The yellow line indicates a baseline where the mask is obtained once and not updated jointly with the tokens (hence fixed value). A clear trend of improvement is observed for all masks. Curves are smoothed to highlight trends.

7.4 **Qualitative results for concept-driven image generation.** Here, we show a comparison between our results with mask optimization, baseline method without mask optimization, and ground truth approach.

7.5 **Comparison of generated masks.** Here we show user-defined masks, masks generated by the baseline method, and those produced by our approach. Note the significant improvement of the mask as a result of joint optimization (first row).
7.6 **Qualitative results demonstrating the interaction between different concepts.** Here we present the interaction between newly learned concepts and the pre-existing concepts already familiar to the stable diffusion model. ................................. 102

7.7 **Diverse images for the same prompt.** Given input concepts, our method demonstrates the ability to generate a variety of diverse outputs for a single prompt. ........................................ 103

7.8 **Interaction between cartoon or real concepts (including background).** Here, our method is not only able to interact with newly learned concepts but also engage with pre-existing concepts (such as Obama and Trump) in a cartoon-like appearance. ................. 103

7.9 **Comparison of our method with standard baselines.** Here we consider TI-m [78], DB-m [270], CD-m [170], ELITE [334], Break-a-scene [16] as standard baseline methods and all images are taken directly from [16]. ....................................................... 104

7.10 **Qualitative results show that our model can handle four concepts simultaneously.** ................................................................. 104

7.11 Qualitative ablation by removing different components. .............. 104

7.12 **Interaction between real-world concepts (including background).** Our model demonstrates the capacity to generate accurate interactions among real-world concepts. ........................................ 105

7.13 **Generated mask w/ and w/o DenseCRF.** ........................................ 106

7.14 **Qualitative results for single concept image generation.** Here, we show a comparison between our method and baseline approach using COCO images. ...................................................... 107

7.15 **Qualitative results for multi-concept image generation.** Here, we show a comparison between our method and baseline approach with interactions where input images are taken from COCO dataset. 108

7.16 **Utilization of multiple images for a single concept.** The model can grasp individual concepts well but faces difficulty in connecting or interacting between these concepts. ................................. 108

7.17 **Stable Diffusion Age Bias.** The model demonstrates a bias towards adult individuals, resulting in a noticeable alteration in the boy’s appearance. ...................................................... 109

7.18 **Stable Diffusion Caucasian Bias.** The model exhibits a stronger bias towards data associated with the Caucasian demographic or region. ...................................................... 109
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D</td>
<td>Two-dimensional</td>
</tr>
<tr>
<td>3D</td>
<td>Three-dimensional</td>
</tr>
<tr>
<td>AI</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>AED</td>
<td>Audio Event Detection</td>
</tr>
<tr>
<td>AG</td>
<td>Attention Gate</td>
</tr>
<tr>
<td>BCE</td>
<td>Binary Cross-Entropy</td>
</tr>
<tr>
<td>BERT</td>
<td>Bidirectional Encoder Representations from Transformers</td>
</tr>
<tr>
<td>BLEU</td>
<td>Bilingual Evaluation Understudy</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional Neural Network</td>
</tr>
<tr>
<td>C3D</td>
<td>Convolutional 3D</td>
</tr>
<tr>
<td>COCO</td>
<td>Common Objects in Context</td>
</tr>
<tr>
<td>CRF</td>
<td>Conditional Random Field</td>
</tr>
<tr>
<td>CV</td>
<td>Computer Vision</td>
</tr>
<tr>
<td>CQT</td>
<td>Constant-Q Transform</td>
</tr>
<tr>
<td>CIDEr</td>
<td>Consensus-based Image Description Evaluation</td>
</tr>
<tr>
<td>CASA</td>
<td>Computational Auditory Scene Analysis</td>
</tr>
<tr>
<td>CT</td>
<td>Computed Tomography</td>
</tr>
<tr>
<td>DEC</td>
<td>Dense Event Captioning</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation-Maximization</td>
</tr>
<tr>
<td>FCN</td>
<td>Fully Convolutional Network</td>
</tr>
<tr>
<td>FM</td>
<td>Foundation Model</td>
</tr>
<tr>
<td>fps</td>
<td>frames per second</td>
</tr>
<tr>
<td>FID</td>
<td>Frechet Inception Distance</td>
</tr>
<tr>
<td>GTX</td>
<td>GeForce GTX series</td>
</tr>
<tr>
<td>GPUs</td>
<td>Graphics Processing Units</td>
</tr>
<tr>
<td>GloVe</td>
<td>Global Vectors for Word Representation</td>
</tr>
<tr>
<td>GPT</td>
<td>Generative Pre-trained Transformer</td>
</tr>
<tr>
<td>GAN</td>
<td>Generative Adversarial Network</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>Hz</td>
<td>Hertz</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>IOU</td>
<td>Intersection over Union</td>
</tr>
<tr>
<td>ICA</td>
<td>Independent Component Analysis</td>
</tr>
<tr>
<td>LDM</td>
<td>Latent Diffusion Model</td>
</tr>
<tr>
<td>LIDAR</td>
<td>Light Detection and Ranging</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Squared Error</td>
</tr>
<tr>
<td>MCB</td>
<td>Multi-modal Compact Bilinear</td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel-Frequency Cepstral Coefficients</td>
</tr>
<tr>
<td>METEOR</td>
<td>Metric for Evaluation of Translation with Explicit ORdering</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic Resonance Imaging</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural language processing</td>
</tr>
<tr>
<td>NMF</td>
<td>Non-negative Matrix Factorization</td>
</tr>
<tr>
<td>NMS</td>
<td>Non-Maximum Suppression</td>
</tr>
<tr>
<td>RGB</td>
<td>Red, Green, and Blue</td>
</tr>
<tr>
<td>RGBD</td>
<td>Red, Green, Blue, and Depth</td>
</tr>
<tr>
<td>R-CNN</td>
<td>Region-based Convolutional Network method</td>
</tr>
<tr>
<td>RNN</td>
<td>Recurrent Neural Network</td>
</tr>
<tr>
<td>ROUGE</td>
<td>Recall-Oriented Understudy for Gisting Evaluation</td>
</tr>
<tr>
<td>ResNet</td>
<td>Residual Network</td>
</tr>
<tr>
<td>RQ</td>
<td>Research Question</td>
</tr>
<tr>
<td>STFT</td>
<td>Short-Time Fourier Transform</td>
</tr>
<tr>
<td>SGD</td>
<td>Stochastic Gradient Descent</td>
</tr>
<tr>
<td>SDR</td>
<td>Signal-to-Distortion Ratio</td>
</tr>
<tr>
<td>SAR</td>
<td>Signal-to-Artifact Ratio</td>
</tr>
<tr>
<td>SIR</td>
<td>Signal-to-Interference Ratio</td>
</tr>
<tr>
<td>SPICE</td>
<td>Semantic Propositional Image Caption Evaluation</td>
</tr>
<tr>
<td>SSS</td>
<td>Sound Source Separation</td>
</tr>
<tr>
<td>SOS</td>
<td>Start of Sequence</td>
</tr>
<tr>
<td>T5</td>
<td>Text-to-Text Transfer Transformer</td>
</tr>
<tr>
<td>URL</td>
<td>Uniform Resource Locator</td>
</tr>
<tr>
<td>ViLBERT</td>
<td>Vision-and-Language BERT</td>
</tr>
<tr>
<td>VAEs</td>
<td>Variational Autoencoders</td>
</tr>
<tr>
<td>VQA</td>
<td>Visual Question Answering</td>
</tr>
<tr>
<td>VCR</td>
<td>Visual Commonsense Reasoning</td>
</tr>
<tr>
<td>VOC</td>
<td>Visual Object Classes</td>
</tr>
</tbody>
</table>
Acknowledgements

First and foremost, I would like to express my deepest gratitude to the Almighty for granting me the capability to successfully finish my entire thesis without any interruptions. His divine wisdom has been my source of strength, inspiration, and perseverance.

I am immensely thankful to my advisor Professor Leonid Sigal for the continuous support of my Ph.D. study and research. Throughout this academic journey, Leonid has not only been a mentor but also a source of inspiration. His expertise, patience, and dedication to the research process have been instrumental in shaping the quality and direction of this work. Leonid provided invaluable feedback, constructive criticism, and encouragement, fostering an environment of intellectual growth and academic excellence. I am truly fortunate to have had the opportunity to work under his supervision.

I extend my heartfelt appreciation to the members of my thesis committee, especially Professor James J. Little, and Professor Kwang Moo Yi. They have been incredibly helpful, offering valuable feedback and suggestions that have made my research much stronger. Their support and insights have been crucial to the quality of my work, and I truly appreciate it.

A special thank you to the University of British Columbia, Department of Computer Science, and Faculty of Graduate and Postdoctoral Studies for providing the necessary resources and facilities that enabled me to conduct my research effectively.

I want to thank my amazing lab friends. They’ve been really nice, helping me a lot and giving great suggestions. And a special thank you to Shweta Mahajan for being an awesome friend and guiding me in a valuable way.

Lastly, I can not forget to give a big shoutout to my wonderful parents and family. Their unconditional love and constant encouragement have been the driving force that brought me to this point and allowed me to pursue and achieve my dream.

Thank you all for being an integral part of this academic endeavor.
Dedication

To my parents, for making me who I am today.
Chapter 1

Introduction

Humans experience their surroundings through multiple modalities of perception by seeing objects, hearing sounds, feeling textures, smelling aromas, and so on. Generally, several senses - visual, auditory, and kinesthetic - help humans to process information for better understanding of the world and remembering more [20, 65, 288]. For example, a silent musical concert is not attractive to the audience. Rather people would like to listen to music/sound and watch performers’ activity in tandem. Similarly, a simple sports newscast with only images/videos for moments of excitement does not provide the full perception of the event; an anchor/commentator with scores (text) running below can greatly enhance the event’s perception. In our daily life, we have many other examples that convey strong evidence that multi-modal data (i.e., vision, audio, text, body language, etc.) helps us to get a complete perception and understanding of a particular event.

Artificial intelligence has primarily focused on unimodal learning to solve tasks related to computer vision, audio, or natural language processing. Researchers have gained tremendous success due to the availability of large datasets (e.g. ImageNet [56], AudioSet [85], BooksCorpus [385]) containing millions of data samples such as images, sounds or words. With the advancement of technology over the last decade, multi-modal content has become avidly available. For example, when we upload a video to social media it contains an audio track along with visual content. Similarly, one can add contextual text in the form of captions or tags with the uploaded images/videos. Live streams, TV shows, and Movies are also accompanied by audio tracks, video, and (sometimes) subtitles. With the abundance, availability, and richness of multi-modal data, learning from only unimodal signals is becoming less natural. Recently, multi-modal learning has made amazing strides in many high-level fundamental visual understanding problems (e.g., image captioning, visual question answering). Learning from multi-modal data not only has been shown to improve the performance of neural networks for specific tasks but also leads to better feature representations useful to make predictions at scale.

Multimodal data has two major benefits: Redundancy and Complementarity [221]. In other words, multimodal learning can capture common/overlapping as well as complementary information from multiple sources, which can enhance predictions compared to unimodal data. Specifically, redundancy helps to improve
performance in the presence of potential unimodal ambiguities; while complementarity allows for a more complete and often detailed understanding of an event. 

Even though sometimes unimodal systems perform better in terms of accuracy and robustness, this can often be attributed to noise and missing concepts in multimodal data [25]. Therefore, the success of multimodal learning depends on many factors including large volumes of multimodal datasets, quality of such data, faster GPUs, high-level feature representations, and so on. The goal of the research in this thesis is to develop and train multimodal neural networks to strengthen correlation and redundancy between multiple modalities to solve high-level image and video understanding problems.

1.1 Motivation and Objectives

Even though multi-modal learning has come a long way and illustrated excellent performance on certain specific tasks (e.g., image captioning, visual question answering), by enlarging it remains an open problem with several unique challenges that stem from the heterogeneity of data involved. We highlight some of the core challenges below:

- **Multi-modal Representations:** Recent works in computer vision and natural language processing (NLP) illustrate the challenge and importance of finding effective representations in their respective modalities of data (visual or lingual). Finding effective joint representations, across multiple modalities, by exploiting both redundancy and complementarity, is even more challenging and remains a core technical challenge in multi-modal learning.

- **Curse of Dimensionality:** Different modalities may have vastly different structures and volumes of data associated with them. For example, images can be represented by high-dimensional tensors capturing pixels, while text may be represented as a sparse vector of word embeddings. Integrating such heterogeneous feature spaces can be computationally expensive and not straightforward, requiring compression and semantic alignment. Further, complete semantic alignment may not even be possible as, for example, some data in one modality may not be represented in another (e.g., certain textures are easy to perceive visually but lack linguistic descriptions).

- **Data Alignment and Annotation:** Learning with aligned and labeled multimodal data is considerably easier, but such data is often difficult to obtain. The data could come unaligned, or labeled at only a certain coarse level (e.g., inherently lacking temporal or spatial alignment). Besides data collection
and annotation generation for all available data is also time-consuming and computationally expensive. This makes multi-modal learning much harder.

- **Fusion of Modalities:** Due to the differences in representations of various modalities (dimensionality and content), effective integration, or fusion, of multi-modal data also remains a significant challenge. For example, visual modality typically has high spatial resolution but low temporal resolution (i.e., visual content changes slowly over time). The auditory signals on the other hand are typically the opposite, i.e., contain low frequency spatial content (due to propagation and reverberation of sound) but very high frequency temporal content. Further, as noted, noise and content itself can differ among modalities; complimentary dictates that certain events will simply not be observed in some of the data modes. These characteristics need to be taken into account explicitly, or implicitly when building multi-modal systems. Therefore, deciding how to fuse or combine information from different modalities is a challenging task. Different fusion techniques may have varying impacts on the model’s performance, and selecting the appropriate fusion strategy is not always straightforward.

- **Model Complexity and Training Challenges:** Multimodal models tend to be more complex than unimodal models. Training such models requires careful consideration of architecture, optimization, and regularization techniques to prevent overfitting and ensure convergence.

- **Evaluation Metrics:** Designing appropriate evaluation metrics for multi-modal tasks is challenging. Traditional metrics may not effectively capture the model’s ability to understand and utilize information from different modalities. For example, there is no appropriate evaluation metric for measuring consistency for the task of multimodal story generation.

- **Computational Intensity:** Processing multiple modalities simultaneously can be computationally intensive. Training and deploying multimodal models may require significant computational resources, limiting their practicality for certain applications. Therefore, efficient use of multimodal data is important.

Moreover, in the real world, one of the modalities could be missing, corrupted, occluded, or mixed with background noise which makes it difficult to train the network or perform inference. Even visual features are significantly affected by small perturbations due to degradation [370]. Besides, data collection and annotation
Chapter 1. Introduction


generation for all available modalities is also time-consuming and computationally expensive. Therefore, to address these challenges, this thesis focuses on the following objectives:

1. Analyzing the importance of different modalities (e.g. audio, human key points) to perform accurate predictions and alignment.

2. Investigating and developing contextual fusion strategies to combine and distill information from multi-modal streams of data.

3. Developing and training multi-modal representations and using them to make predictions from human-centric videos (containing actors, actions, and gestures).

4. Further, we consider weak supervision to overcome limitations of missing and limited data. Specifically, instead of relying on full spatial-temporal annotations, we use partially labeled data to train the developed networks.

5. Finally, we explore multi-modal generative models to enable creativity and creative exploration, by focusing on image/video generation conditioned on text. In doing so, we focus on the consistency of generated visuals.

6. While text-to-image models demonstrate strong performance, they strongly rely on large-scale training. Therefore, we also explore efficient personalization through subject-driven generative model adaptation.

1.2 Thesis Outline and Contributions

In this section, we briefly summarise the contributions and provide an outline of each chapter. Given the diversity of topics covered, the thesis contributions are divided into six chapters (Chapters 3 to 8) with corresponding backgrounds and related works in each. The main contributions of the thesis are outlined as follows.

Multimodality in Visual-lingual Tasks: As the first step, we explored the importance of multimodal data in the visual-lingual domain. Specifically, we focused on weakly supervised dense video captioning in videos. Dense video captioning is a challenging task that entails the identification of multiple events within a video and providing descriptive narratives for each of these events using natural language. In our research, we enhance state-of-the-art performance by merging audio and visual features. We believe that sound might play a more crucial role than is commonly acknowledged in this context. To the best of our knowledge, ours is the first work that addressed dense event captioning by leveraging auditory data.
The work has two key contributions: we (1) show the importance of audio features, and (2) introduce different fusion strategies to fuse multimodal (visual and auditory) embeddings. We also tackle dense event captioning in a weakly supervised manner, not requiring annotations of temporal video segments in training. Merging audio and visual representations is critical for the task, therefore we discuss and test different fusion strategies (e.g. Multiplicative mixture fusion, Multimodal context fusion, and MUTAN fusion) for incorporating audio and visual cues. Chapter 3: Multi-modal Weakly Supervised Dense Event Captioning discusses the developed approach in depth and presents corresponding results. The work was first published at ICCV in 2019 [258].

Sound Source Detection and Separation: In a complex auditory scene, audio could be mixed with background “noise” and isolating individual sound sources from the mixture is quite challenging. Therefore, as the second step, we consider self-supervised learning and focus on learning how to localize and separate individual object sounds within the audio channel of the video. We propose an audio-visual co-segmentation, where the network learns both what individual objects look and sound like, from videos labeled with only object labels. Our architecture can be learned in an end-to-end manner and requires no additional supervision or bounding box proposals. More generally, we introduce weakly-supervised object segmentation in the context of sound separation and formulate spectrogram mask prediction using a set of learned mask bases. Later we combine these bases with coefficients conditioned on the output of object segmentation. Chapter 4: Weakly-supervised Audio-visual Sound Source Detection and Separation discusses the approach in detail and presents exhaustive experimental results. The work was published at ICME in 2021 as an oral presentation [256].

Multimodal Human-centric Generic Representation Learning: The ultimate goal of our research is to learn aligned representations to find how humans perceive multi-sensory data. The quality of these representations has been shown to greatly impact the overall performance of downstream tasks. Therefore, in Chapter 5: Human-centric Audio-visual Representation Learning, we introduce TriBERT – a transformer-based architecture, inspired by ViLBERT, which can learn generic feature representations across three modalities: vision, pose, and audio, with the use of flexible co-attention (i.e. Tri-modal co-attention) layer. From a technical perspective, as part of the TriBERT architecture, we introduce a learned visual tokenization scheme based on spatial attention and leverage weak supervision to allow granular cross-modal interactions for visual and pose modalities. Further, we supplement learning with sound-source separation loss formulated across all three streams. The work was published at NeurIPS 2021 [304].
Multimodality in Generative AI: Generative AI has emerged as one of the most influential areas recently. In this thesis, we specifically set forward to explore text-to-image generative models that are designed to generate images conditioned on text. Specifically, we focused on story generation, the goal of which is to generate a sequence of illustrative image frames with coherent semantics given a sequence of sentences. In our work, for the first time (to our knowledge), we study coreference resolution in story generation. That is, we resolve references (i.e. he/she/they) instead of requiring actors to be named explicitly each time. Moreover, we ensure stories are illustrated consistently with subject and background appearance not changing between frames (unless explicitly required by the story text). We introduce Story-LDM [257], a deep generative approach with autoregressive structure and an extension of Latent Diffusion Model [266]. We propose a novel memory attention mechanism that takes into account the already generated semantics of the previous frames to ensure temporal consistency and smooth story progression. To validate co-reference resolution, and character and background consistency, we extend existing datasets and evaluation metrics to include more complex scenarios. Chapter 6: Visual Memory Conditioned Consistent Story Generation provides a detailed explanation of the methodology employed and presents the outcomes of the experiments. The work was published at CVPR in 2023 [257].

Generative AI Personalization: The rise of generative AI has led to remarkable achievements in text-to-image models, enabling the creation of high-quality and diverse synthesized images based on textual prompts [259, 266, 272]. These models typically draw on semantic patterns learned from pairs of images and captions. However, a limitation arises in their capability to generate specific subjects—such as objects, animals, or individuals—with novel and distinct characteristics. To address this challenge, innovative approaches have emerged to personalize text-to-image diffusion models [170, 270]. These methods involve refining pre-trained text-to-image models through fine-tuning, enabling them to acquire unique identifiers associated with specific subjects within various contextualized scenes. But what if individuals wish to introduce or replace dream characters within existing images? What if they aim to maintain the characters but alter the settings? This leads us to envision a more tailored output generation process – one that places the desired character seamlessly into a particular world. Examples include scenarios like “Mickey engaging in combat with Batman within the DC universe” or “Cinderella conversing with Fred in the imaginative Flintstone world”. Therefore, in 5th step, we proposed concept-driven image generation. In Chapter 7: Visual Concept-driven Image Generation with Text-to-Image Diffusion Model, we thoroughly explain the methods we used in this work and present the experimental outcomes in a more detailed manner. The work is under review at CVPR 2024.
Finally, in Chapter 8: Conclusions, we explore the implications of our research on computer vision and the machine learning community. It serves as a critical reflection on the potential applications and advancements our research could bring to these fields.
1.3 Publications

Each chapter from 3 to 7 represents a paper that has been peer-reviewed, accepted, or submitted to a conference. Therefore, the publications included in this thesis are given below:

Chapter 3: “Watch, Listen and Tell: Multi-modal Weakly Supervised Dense Event Captioning”

Chapter 4: “Weakly-supervised Audio-visual Sound Source Detection and Separation”

Chapter 5: “TriBERT: Human-centric Audio-visual Representation Learning”

Chapter 6: “Make-A-Story: Visual Memory Conditioned Consistent Story Generation”

Chapter 7: “Visual Concept Driven Image Generation with Text-to-Image Diffusion Model”
Publications not included:

“An Improved Attention for Visual Question Answering”
Tanzila Rahman, Shih-Han Chou, Leonid Sigal and Giuseppe Carenini. 4th Multimodal Learning and Applications Workshop (in conjunction with CVPR), 2021.

“Prompting Hard or Hardly Prompting: Prompt Inversion for Text-to-Image Diffusion Models”
Chapter 2

Background

This thesis explores efficient learning from multiple modalities in service of perception and generation of visual content. In this chapter, we present foundational aspects and an overview of related works in multimodal learning. We also discuss the fundamental principles and practical uses of multimodal learning.

2.1 Foundations of Multimodal Learning

Multimodal machine learning is a dynamic and interdisciplinary research field at the intersection of computer science, artificial intelligence, and cognitive science. It seeks to empower computer systems with the ability to comprehend, reason, and learn from a diverse range of perceptual and communicative signals, including but not limited to linguistic, acoustic, visual, tactile, and physiological modalities. This comprehensive approach aims to create more intelligent and context-aware systems capable of understanding and communicating with humans.

Before the widespread adoption of deep learning methods, approaches for multimodal learning, especially tasks involving both vision and language, relied on traditional machine learning techniques, statistical models, and handcrafted features. Specific features were manually extracted from images using methods like SIFT [201], and from text using techniques such as bag-of-words [366]. Hidden Markov Models (HMMs) [245] used for sequential data modeling and often applied in speech recognition and part-of-speech tagging tasks. The Naive Bayes Classifier [24], based on Bayes’ theorem [142], was employed to assume a strong dependency between features. Researchers also utilized maximum entropy models [212] to represent the probability distribution of features given a label.

Support Vector Machine (SVM) [243] was a widely used machine learning algorithm for tasks like image and text classification. It operated by mapping input data to a high-dimensional feature space and identifying the optimal hyperplane to separate different classes. Kojima *et al.* [157] introduced a method that generated natural language descriptions of video clips by combining visual and linguistic features with a probabilistic language model. In another study, Krishnamoorthy *et al.* [167] aimed to enhance automated video understanding by incorporating ex-
ternal textual information to produce coherent and informative descriptions, thus improving the interpretability and usefulness of video content analysis systems. Farhadi et al. [70] introduced a method for generating textual descriptions of images by mapping visual features extracted from images to textual concepts using a probabilistic graphical model. Additionally, Ordonez et al. [229] proposed a method for generating image descriptions using a multimodal topic model that integrates visual and textual features. Hodosh et al. [116] proposed a method for training models to generate descriptions by ranking candidate sentences based on their relevance to the input image, facilitating more accurate and interpretable image captioning systems. An approach involving extracting features from both the image and the associated news article was proposed in [73], which was then used to train a caption generation model. The resulting captions were evaluated for their relevance, coherence, and informativeness, with promising results indicating the potential of the approach to enhance news content understanding and accessibility. Furthermore, researchers have shifted their focus toward spatio-temporal features to tackle a variety of computer vision tasks. Laptev et al. [172] introduced space-time interest points (STIPs) as local features for spatio-temporal video analysis. STIPs are detected based on gradients of motion and appearance and have been widely used for action recognition tasks. Spatio-temporal descriptors based on histograms of oriented gradients (HOG) [51] and histograms of optical flow (HOF) [241] for action and gesture recognition were introduced in [274]. Wang et al. [325] presented an improved trajectory-based representation for action recognition, which incorporates both spatial and temporal information. It uses features such as trajectory shape, motion, and context to encode information in videos.

Today, machine learning, particularly deep learning, and computer vision are placing greater emphasis on multimodal learning to attain perception levels akin to humans. It involves integrating information from various sources, such as text, images, and audio, to enhance understanding. However, the development of this approach is not an overnight success but rather the result of extensive prior research and groundwork.

Lingual: Groundbreaking work in linguistic modality includes the development of vector word embeddings. The work of Mikolov et al. [216] on Word2Vec and subsequent advancements, like GloVe [240], have revolutionized natural language processing (NLP) by providing effective ways to represent and understand textual information. The subsequent advent of sequence-to-sequence models [299] and attention mechanisms [317] marked a transformative period. Recent advancements in NLP, such as BERT [58], GPT [249], and T5 [251], have further pushed the boundaries of natural language understanding, enabling machines to capture intricate contextual relationships in textual data.
**Visual:** In the visual domain, the introduction of deep convolutional neural networks (CNNs), as demonstrated by Krizhevsky *et al.* [168], has significantly advanced image classification. Object detection has seen notable progress with the development of region-based CNNs (R-CNN) [88] and its iterations, such as Faster R-CNN [87] and Mask R-CNN [106]. Fully Convolutional Networks (FCN) [200] and U-Net [267] are two influential architectures in the field of semantic segmentation. They have set the foundation for subsequent advancements in semantic segmentation, and researchers continue to build upon these architectures to improve performance and address specific challenges in different domains. In the realm of generative models, Goodfellow *et al.* introduced the concept of GANs [89], where a generator and discriminator are trained in tandem, inspiring subsequent advancements in the field. Building upon this foundation, Mirza *et al.* extended GANs to a conditional setting [218], enabling the generation of specific data based on given conditions. In [152], Kingma *et al.* introduced Variational Autoencoders (VAEs), which learn structured representations of data and have become influential in various generative tasks. Additionally, [34, 144] focused on producing high-fidelity natural images with improved quality and diversity.

**Auditory:** Researchers have recently uncovered the potential for significant information within audio that was previously overlooked by the community. Wav2vec 2.0 [19] demonstrates a groundbreaking finding - if we can extract useful speech details from unlabeled data, it can enhance how well we do in tasks related to speech later on. Auto-encoding is a common method used to recreate signals, like speech. It involves two steps: first, converting the speech into a compact representation, and then turning that representation back into speech [164]. Methods of self-supervised learning (SSL) in speech have achieved impressive results in tasks like automatic speech recognition [276, 324], phoneme segmentation [163], and audio compression [228, 357]. Later, Hsu *et al.* [123] introduced the HuBERT model. It is trained using a masked prediction task, similar to BERT, but instead of text, it works with masked continuous audio signals.

In recent years, the intersection of computer vision and audio modality has witnessed notable developments, expanding the scope of applications to encompass multimodal understanding. Therefore, researchers often use both sound and images together to make things work better than when using just one of them. Current audio-visual learning tasks could fall into many sub-fields including audio-visual separation and localization [246, 339, 352, 381], generation [192, 195, 328, 384], correspondence [5, 166, 217, 219] and representation learning [45, 275, 304, 316].

Early multimodal approaches have focused on modal-specific encoder (e.g., CNNs for images) – decoder (e.g., RNNs for text) architectures, trained jointly for
specific tasks on paired data. This has, however, evolved with current research largely focusing on the Foundation Models (FM) trained either across or for broad variety of tasks, often in unsupervised or self-supervised manner. FMs have been shown to be really good at understanding, translating, and creating different types of information across various modes. In multimodal FM research, there are two main approaches [345]. One involves putting data from different sources into a common latent space, usually using transformer encoders. The other approach is about creating data in various forms, often using transformer decoders. The early FM works, CLIP [247], ALBEF [181], and ALIGN [137], were pioneers in introducing cross-modal alignment as multi-encoder FMs. They aimed to improve representations by establishing connections between images and text. Later works like ImageBind [86] and LanguageBind [375] used different intermediate modalities as a common ground, effectively mapping representations from different sources into this shared space. This made it easier to transform information across different modes within a unified vector space. In addition, works like BLIP-2 [180] and MiniGPT-4 [376] concentrate on efficient learning by utilizing frozen image encoder and language decoder, pre-trained on uni-modal data, and learning alignment (in the form of Q-Former – a lightweight transformer layer) between them. Meanwhile, mPLUG [351] and LLaVA [193] work on improving the ability to transform different types of information, enhancing the reliability of the generated outcomes. Additionally, Flamingo [7] and LLaMA-Adapter [363] explore ways to adjust multimodal large FMs more efficiently, in the case of Flamingo through in-context learning, resulting in higher-quality multimodal outputs with lower cost.

Stable diffusion [265] can generate high-quality images, by adding noise and then removing it through a learned method. This model is used for different downstream tasks like text-to-image generation, image inpainting, image editing, image super-resolution, and so on. Besides stable diffusion, there is a significant focus on “any-to-any” generative models (e.g. CoDi [302], NExT-GPT [336]) that can transform various types of inputs into a wide range of outputs such as language, image, video, or audio. Following works like FLAVA [287], CoCa [356], and GLIP [182] go deeper into improving how decoders can effectively combine and align representations of both images and text. This enhances the ability to reason across different modes. Moreover, SAM [155] takes it a step further by using a decoder to merge prompt embeddings related to images, allowing for the automatic segmentation of images based solely on text prompts, even without prior training.

Nowadays researchers have contributed a lot by proposing methods using multimodal data to solve many computer vision tasks such as multi-modal image/video captioning [222, 305, 338, 348], speech synthesis [151, 273, 354], emotion recognition [4, 160], cross-modal retrieval [48, 80, 360] and many more. In this chapter,
we cover only the basic concepts and core advances related to the diverse field of multimodal learning. A more thorough discussion and taxonomy of prior works can be found in numerous surveys, most notably [21, 26, 95, 188, 345, 346, 377].

2.2 Core Principles in Multimodal Research

Multimodal research investigates different types of information, such as text, images, sounds, and videos. The core principles of this research focus on understanding how these different pieces of information connect and work together. For example, figuring out how text matches with images or how sounds relate to videos. The core principles underscore the interconnected and heterogeneous nature of these modalities, emphasizing the need to study them collectively rather than in isolation. Therefore, from a research perspective, the study of multimodal information can be broken down into three fundamental principles [188]:

2.2.1 Modalities are Diverse

When we say that modalities are heterogeneous, we mean that the information found in different modalities tends to be diverse in terms of its characteristics, structure, and representations. Let’s break down this idea into more detail:

- **Diverse qualities:** Each modality, whether it’s text, images, audio, or video, carries information in a distinct manner. For example, text conveys information through words and language, while images capture visual details, and audio provides information through sounds. The qualities of these modalities vary significantly due to their unique nature.

- **Structural Differences:** The way information is organized and presented within each modality can differ. Text has sentences and paragraphs, images have pixels arranged in a specific pattern, audio has waveforms, and videos consist of a sequence of frames. These structural differences make it challenging to directly compare or analyze information across modalities without considering their specific structure.

- **Representational Variations:** Information representation refers to how data is encoded or expressed within a modality. For instance, text can be represented as a sequence of words, images as arrays of pixel values, and audio as waveforms. These variations in representation further contribute to the heterogeneity of modalities, requiring specialized methods to interpret and process the information effectively.
Understanding and acknowledging the heterogeneous nature of modalities is crucial in multimodal research. Researchers develop methods and models that can handle this diversity, allowing for the integration and analysis of information from different modalities.

### 2.2.2 Modalities are Interconnected

Even though different types of information (modalities) can be quite diverse in terms of content, structure, and representations, they are often connected because they share helpful information that complements each other. They can be considered as different views of the same underlying scene or event. Consider a scenario where we have both a written description (text) and an accompanying image of a scene. The words may describe the sunny weather and a beach, while the image shows people enjoying themselves in the sunlight. The words and the image, although different modalities, share complementary information about the scene, forming a connection that enhances our understanding.

In investigating these connections, researchers employ statistical approaches [308]. For instance, statistical association involves recognizing how often specific words in the text co-occur with certain objects or scenes in the image. If the word “beach” often appears with images of people in swimwear, there is a statistical association. Going deeper, statistical dependence [225, 313] explores whether the presence of certain words in the text causally relates to the appearance of specific objects in the image. On the semantic side, researchers may focus on semantic correspondence [230]. In our example, it involves determining which words in the text correspond to elements in the image. If the word “sunny” in the text corresponds to the depiction of bright sunlight in the image, it demonstrates semantic correspondence. Semantic relations take this a step further by considering attributes that describe the nature of the relationship [109, 214]. For instance, not just recognizing the shared meaning of “sunny” but understanding the specific attribute of brightness it brings to the scene.

### 2.2.3 Modalities Interact

Modality interactions explore how different types of information work together to create new insights when used for a task. It’s important to understand the difference between connections and interactions: connections are relationships within the information itself, while interactions happen when we combine and process these different types to get a new result. The dimensions of interactions involve looking at the type of connected information (whether it’s shared or unique), understanding the ways we combine them (like adding them up or using logical operations),
and studying how our response changes when we use multiple types of information (whether it is the same, better, or different). This helps us see how integrated information contributes to our understanding of tasks.

When we consider multimodal interaction, *complementarity* and *redundancy* play pivotal roles between different modes of information [221]. Complementarity is different ways of presenting information (*e.g.* images, text, and sound) each providing unique details that, when combined, give a fuller picture. For example, imagine learning about animals using a website that has both pictures and written descriptions. The pictures show how they look, while the text gives more details about their behavior. Together, they complement each other, making our learning experience richer.

On the flip side, redundancy in modalities interaction is like having backup information. It is when different ways of presenting information share some details, making sure we understand even if one part is not clear. For instance, in a cooking video, the chef might explain a recipe both by talking and showing the steps. If we miss something in the audio, we can still get it from the visuals, or vice versa. This redundancy ensures we catch all the important details.

### 2.3 Applications of Multimodal Learning:

Multimodal learning finds diverse applications across various domains, harnessing the synergistic power of multiple modalities to enhance intelligent systems. Below are key applications with references highlighting significant contributions and advancements.

- **Human-Computer Interaction (HCI):** Multimodal learning plays a pivotal role in creating more natural and intuitive interactions between humans and computers. Gesture recognition, voice commands, and sentiment analysis contribute to a seamless user experience [18].

- **Healthcare:** In healthcare, multimodal learning enables the analysis of various physiological signals for improved diagnostics, patient monitoring, and personalized treatment plans. Integration of diverse modalities enhances the accuracy and efficiency of healthcare systems [103].

- **Autonomous Systems:** Autonomous systems, particularly in the context of self-driving cars, leverage multimodal learning for enhanced perception and decision-making [42]. Combining information from sensors such as cameras, LIDAR, and radar improves the system’s understanding of its environment.
• **Cross-Modal Transfer Learning:** The exploration of cross-modal transfer learning facilitates the transfer of knowledge learned from one modality to another, addressing data scarcity challenges and improving model generalization [76].

• **Education:** Multimodal learning contributes to personalized and adaptive learning systems. By incorporating various modalities such as text, images, and speech, educational platforms can tailor content delivery to individual learning styles.

• **Security and Surveillance:** In security and surveillance applications, multimodal learning enhances the capabilities of video analysis systems [281]. Integration of visual and acoustic modalities allows for more robust threat detection and activity recognition.

• **Virtual and Augmented Reality (VR/AR):** Multimodal learning is crucial in creating immersive experiences in virtual and augmented reality environments. By combining visual, auditory, and haptic modalities, these systems offer a more realistic and interactive user experience [6].

Multimodal learning has a big impact on technology and how machines interact with humans. Researchers are always finding new ways and methods to make this field even more exciting. This interdisciplinary area holds the promise of groundbreaking discoveries that could change how machines see and understand the world.
Chapter 3

Multi-modal Weakly Supervised Dense Event Captioning

Humans often perceive the world through multiple sensory modalities, such as watching, listening, smelling, touching, and tasting. Consider two people sitting in a restaurant; seeing them across the table suggests that they may be friends or coincidental companions; hearing, even the coarse demeanor of their conversation, makes the nature of their relationship much clearer.

In our daily lives, many other examples produce strong evidence that multi-modal co-occurrences give us a fuller perception of events. Recall how difficult it is to perceive the intricacies of the story from a silent film. Multi-modal perception has been widely studied in areas like psychology [53, 323], neurology [292], and human-computer interaction [303].

In the computer vision community, however, the progress in learning representations from multiple modalities has been limited, especially for high-level perceptual tasks where such modalities (e.g. audio or sound) can play an integral role. Recent works [231, 279] propose approaches for localizing audio in unconstrained videos (sound source localization) or utilizing sound in video captioning [101, 120, 306, 331]. However, these approaches consider relatively short videos, i.e., usually about 20 seconds, and focus on the description of a single salient event [344]. More importantly, while they show that audio can boost the performance of visual models to an extent, such improvements are typically considered marginal and the role of audio is delegated to being secondary (or not nearly as important) as visual signal [120, 331].

We posit that sound (or audio) may be much more important than the community may realize. Consider the previously mentioned example of a silent film. The lack of sound makes it significantly more difficult, if not impossible in many cases, to describe the rich flow of the story and constituent events. Armed with this intuition, we focus on dense event captioning [184, 327, 372] (a.k.a. dense-captioning of events in videos [165]) and endow our models with the ability to utilize rich auditory signals for both event localization and captioning. Figure 3.1 illustrates one example of our multi-modal dense event captioning task. Compared with conventional video captioning, dense event captioning deals with longer and
A group of friends sit together in a living room talking. One of the friends plays a song on his guitar and the others listen. One friend falls back into the chair in amazement after the song. The two friends congratulate the man for his guitar performance and shake hands.

Figure 3.1: **Multi-modal Dense Event Captioning.** Illustration of our problem definition, where we use both audio features and visual information to generate the dense captions for a video in a weakly supervised manner.

more complex video sequences, usually 2 minutes or more. To the best of our knowledge, our work is the first to tackle dense event captioning with sound, treating sound as a first-class perceptual modality.

Audio features can be represented in many different ways. Choosing the most appropriate representation for our task is challenging. To this end, we compare different audio feature representations in this work. Importantly, we show that audio signal alone can achieve impressive performance on the dense event captioning task (rivaling visual counterpart). The form of fusion needed to incorporate the audio with the video signal is another challenge. We consider and compare a variety of fusion strategies.

Dense event captioning provides detailed descriptions for videos, which is beneficial for in-depth video analysis. However, training a fully supervised model requires both caption annotations and corresponding temporal segment coordinates (*i.e.*, the start and end time of each event), which is extremely difficult and time-consuming to collect. Recently, [63] proposed a method for dense event captioning in a weakly supervised setting. The approach does not require temporal segment annotation during training. During evaluation, the model can detect all events of interest and generate their corresponding captions. Inspired by and building on [63], we tackle our multi-modal dense event captioning in a weakly supervised manner.

**Contributions.** Our contributions are multiplefold. First, to the best of our knowl-
Chapter 3. Multi-modal Weakly Supervised Dense Event Captioning

20

edge, this is the first work that addresses dense event captioning tasks in a multi-modal setting. In doing so, we propose an attention-based multi-modal fusion model to integrate both audio and video information. Second, we compare different audio feature extraction techniques [17, 54, 189], and analyze their suitability for the task. Third, we discuss and test different fusion strategies for incorporating audio cues with visual features. Finally, extensive experiments on the ActivityNet Captions dataset [165] show that the audio model, on its own, can nearly rival the performance of a visual model and, combined with video, using our multi-modal weakly-supervised approach, can improve on the state-of-the-art performance.

3.1 Related Work

Audio Feature Representations. Recently computer vision community has begun to explore audio features for learning good representations in unconstrained videos. Aytar et al. [17] propose a sound network guided by a visual teacher to learn the representations of sound. Earlier works, [231, 279, 296], address sound source localization problem to identify which pixels or regions are responsible for generating a specified sound in videos (sound grounding). For example, [279] introduces an attention-based localization network guided by sound information. A joint representation between audio and visual networks is presented in [231, 296] to localize sound sources. Gao et al. [82] formulate a new problem of audio source separation using a multi-instance multi-label learning framework. This framework maps audio bases, extracted by non-negative matrix factorization (NMF), to the detected visual objects. In recent years, audio event detection (AED) [36, 236, 301] has received attention in the research community. Most of the AED methods locate audio events and then classify each event.

Multi-modal Features in Video Analysis. Combining audio with visual features (i.e., multi-modal representation) often boosts the performance of networks in vision, especially in video analysis [12, 15, 120, 306, 331]. Ariav et al. [15] propose an end-to-end deep neural network to detect voice activity by incorporating audio and visual modalities. Features from both modalities are fused using multi-modal compact bilinear pooling (MCB) to generate a joint representation for speech signals. Authors in [12] propose a multi-modal method for egocentric activity recognition where audio-visual features are combined with multi-kernel learning and boosting.

Recently, multi-modal approaches are also gaining popularity for video captioning [306, 331]. In [120] a multi-modal attention mechanism to fuse information across different modalities is proposed. Hori et al. [121] extend the work
in [120] by applying hypothesis-level integration based on minimum Bayes-risk decoding [169, 293] to improve the caption quality. Hao et al. [101] present multi-modal feature fusion strategies to maximize the benefits of visual-audio resonance information. Wang et al. [331] introduce a hierarchical encoder-decoder network to adaptively learn the attentive representations of multiple modalities, and fuse both global and local contexts of each modality for video understanding and sentence generation. A module for exploring modality selection during sentence generation is proposed in [306] to interpret how words in the generated sentences are associated with audio and visual modalities.

**Dense Event Captioning in Videos.** The task of dense event captioning in videos was first introduced in [165]. The task involves detecting multiple events that occur in a video and describing each event using natural language. Most of the works [196, 350] solve this problem in a two-stage manner, i.e., first temporal event proposal generation and then sentence captioning for each of the proposed event segments. In [350], authors adopt a temporal action proposal network to localize proposals of interest in videos, and then generate descriptions for each proposal. Wang et al. [327] present a bidirectional proposal method that effectively exploits both past and future contexts to make proposal predictions. In [372], a differentiable masking scheme is used to ensure the consistency between proposal and captioning modules. Li et al. [184] propose a descriptiveness regression component to unify the event localization and sentence generation. Xu et al. [343] present an end-to-end joint event detection and description network (JEDDi-Net) which adopts region convolutional 3D network [342] for proposal generation and refinement, and proposes hierarchical captioning.

Duan et al. [63] formulate the dense event captioning task in a weakly supervised setting, where there are no ground-truth temporal segment annotations during training and evaluation. They decompose the task into a pair of dual problems, event captioning, and sentence localization, and present an iterative approach for training. Our work is motivated by [63] and builds on their framework. However, importantly, we fuse audio and visual features and explore a variety of fusion mechanisms to address the multi-modal weakly supervised dense event captioning task. We note that [63] is thus far the only method for dense event captioning in the weakly supervised setting.

### 3.2 Multi-modal Dense Event Captioning

In this work, we consider two important modalities, audio and video, to generate dense captions in a weakly supervised setting. Weak supervision means that we do
Our Multi-modal Architecture. The model has two parts, a sentence localizer and a caption generator. The sentence localizer takes audio, video, and captions as inputs and generates a temporal segment for each caption. The caption generator uses the resultant temporal segments, with audio and video features, to produce a caption for each segment.

not require ground-truth temporal event segments during training. The overview of our multi-modal architecture is shown in Figure 3.2. The architecture consists of two modules, a sentence localizer and a caption generator. Given a set of initial random proposal segments in a video, a caption generator produces captions for the specified segments. The sentence localizer then refines the corresponding segments with the generated captions. The caption generator is employed again to refine the captions. This process can proceed iteratively to arrive at consistent segments and captions; in practice, we use one iteration following the observations in [63].

We extract features from audio, video, and captions first, and pass them as inputs to the sentence localizer during training. For each modality, an encoder is used to encode the input. We use recurrent neural networks (RNNs) with GRU [47] units as encoders. We then apply a crossing attention among the audio, video, and caption features. Then an attention feature fusion mechanism followed by a fully connected layer is applied to produce temporal segments.

The caption generator takes the encoded features of audio and video, along with the resultant temporal segments as inputs. It performs soft mask clipping on the audio and video features based on the temporal segments, and uses a context fusion technique to generate the multi-modal context features. Then a caption decoder, which is also an RNN with GRU units, generates one caption for each multi-modal context feature. We discuss and compare three different context fusion strategies to find the most appropriate one for our multi-modal integration.

In what follows, we first describe how to extract features from audio and video in Sec. 3.2.1. Then we present our weakly supervised approach in Sec. 3.2.2.
Lastly, we demonstrate three different context fusion strategies in Section 3.2.3.

### 3.2.1 Feature Representation

We consider both features from audio and video modalities for dense event captioning. It is generally challenging to select the most appropriate feature extraction process, especially for the audio modality. We describe different feature extraction methods to process both audio and video inputs.

#### Audio Feature Processing

ActivityNet Captions dataset [165] does not provide audio tracks. As such, we collected all audio data from the YouTube videos via the original URLs. Some videos are no longer available on YouTube. In total, we were able to collect around 15,840 audio tracks corresponding to ActivityNet videos. To process the audio, we consider and compare three different audio feature representations.

**MFCC Features.** Mel-Frequency Cepstrum (MFC) is a common representation of sound in digital signal processing. Mel-Frequency Cepstral Coefficients (MFCCs) are coefficients that collectively make up an MFC – a representation of the short-term power spectrum of sound [138]. We down-sample the audio from 44 kHz to 16 kHz and use 25 as the sampling rate. We choose 128 MFCC features, with 2048 as the FFT window size and 512 as the number of samples between successive frames (i.e., hop length).

**CQT Features.** The Constant-Q-Transform (CQT) is a time-frequency representation where the frequency bins are geometrically spaced and the ratios of the center frequencies to bandwidths (Q-factors) of all bins are equal [35]. CQT is motivated by the human auditory system and the fundamental frequencies of the tones in Western music [277]. We perform feature extraction by choosing 64 Hz and 60 as the minimum frequency and the number of frequency bins respectively. Similar to the MFCC features described above, we use 2048 as the FFT window size and 512 as the hop length. We use VGG-16 [285] without the last classification layer to convert both MFCC and CQT features into 512-dimensional representations.

**SoundNet Features.** SoundNet [17] is a CNN that learns to represent raw audio waveforms. The acoustic representation is learned using two million videos with their accompanying audio; leveraging the natural synchronization between them. We use a pre-trained SoundNet [17] model to extract the 1000-dimension audio features from the 8-th convolutional layer (i.e., conv8) for each video’s audio track.
Chapter 3. Multi-modal Weakly Supervised Dense Event Captioning

Video Feature Processing

Given an input video $V = \{v_t\}_{t=1}^{T_v}$, where $v_t$ is the video frame at time $t$ and $T_v$ is the video length, a 3D-CNN model is used to process the input video frames into a sequence of visual features $\{f_t = F(v_t : v_{t+\delta})\}_{t=1}^{T_f}$. Here, $\delta$ means the time resolution for each feature $f_t$ and $T_f$ is the length of the feature sequence. We use features extracted from encoder $F$ provided by the ActivityNet Captions dataset [165], where $F$ is the pre-trained C3D [136] network with $\delta = 16$ frames. The dimension of the resultant C3D features is a tensor of size $T_f \times D$, where $D = 500$ and $T_f = T_v / \delta$.

3.2.2 Weakly Supervised Model

Weak supervision means that we do not require ground-truth temporal alignments between the video (visual and audio collectively) and captions. We make a one-to-one correspondence assumption, meaning that we assume that each caption describes one temporal segment and each temporal segment corresponds to only one caption. This assumption holds in the current benchmark dataset and most real-world scenarios. We employ two network modules, a sentence localizer and a caption generator. Given a caption, the sentence localizer will produce a temporal segment in the context, while the caption generator will generate a caption with a given temporal segment. We use context to refer to an encoded video or audio.

Notations. We use GRU RNNs to encode visual and audio streams of the video. This results in a sequence of output feature vectors, one per frame, $O = \{o_t \in \mathbb{R}^k\}_{t=0}^{T_o}$ and the final hidden state $h_o \in \mathbb{R}^k$, where $T_o$ is the length of the video. While in practice we get two sets of such vectors (one set for video and one set for corresponding audio “frames”), we omit the subscript for clarity of formulation.
that follows. A caption is encoded similarly by the output features of the RNN, $C = \{c_t \in \mathbb{R}^k\}_{t=0}^{T_c}$ with the last hidden state being $h^c \in \mathbb{R}^k$, where $T_c$ is the length of the caption in words. We use context to refer the encoding of the full visual or audio information in videos. A context segment $S$ is represented by $(c, l)$, where $c$ and $l$ denote the segment’s temporal center and length respectively within $O$.

Sentence Localizer

Sentence localizer attempts to localize a given caption in a video by considering the caption and the encoded complete video (context). Formally, given a (video or audio) context $O$ and an encoded caption $C$, sentence localizer will regress a temporal segment $S$ in $O$. With the context and caption features, it first applies crossing attention among them. Then an attention feature fusion, followed by one layer fully-connected neural network, is used to generate the temporal segment. Following [63], we use 15 predefined temporal segments and generate 15 offsets in sentence localization using a fully connected layer. The final segments are the sum of temporal segments and offset value. The purpose is to fine-tune the offset value for best localization.

Crossing Attention. The crossing attention consists of two sub-attentions, one caption attention $\text{Att}_{c}$, and one context attention $\text{Att}_{o}$. For a context $O$ and a caption $C$, we first compute the attention between $h^o$ and $C$ as:

$$\text{Att}_{c} = \text{softmax}((h^o)^T \alpha_c C)C^T,$$

and then calculate the attention between $h^c$ and $O$ as:

$$\text{Att}_{o} = \text{softmax}((h^c)^T \alpha_o O)O^T,$$

where $\alpha_c \in \mathbb{R}^{k \times k}$ and $\alpha_o \in \mathbb{R}^{k \times k}$ are the learnable attention weights, and $(\cdot)^T$ is the matrix transpose operation. We note that $\text{Att}_{o}$ is a vector of size $1 \times k$ comprising of attention-weighted features for the visual/audio frames; similarly $\text{Att}_{c}$ is a vector of size $1 \times k$ of attended caption features.

When training our multi-modal approaches, the caption attention $\text{Att}_{c}$ is calculated only between the visual modality and the captions, and we generate video attention $\text{Att}_{v}$ and audio attention $\text{Att}_{a}$ using Eq. 3.2. While we are training our unimodal approaches which either use audio (or video) information to generate captions, the caption attention $\text{Att}_{c}$ is calculated between the audio (or video) and captions.
Attention Feature Fusion. After obtaining the sub-attentions, we use the multi-model feature fusion technique [81] to fuse them together:

\[
\begin{align*}
\text{Att}_{\text{sum}} &= \text{Att}_c + \text{Att}_v + \text{Att}_a \\
\text{Att}_{\text{dot}} &= \text{Att}_c \cdot \text{Att}_v \cdot \text{Att}_a \\
\text{Att}_{\text{fc}} &= \text{fc}(\text{Att}_c||\text{Att}_v||\text{Att}_a) \\
\text{Att}_{\text{fusion}} &= \text{Att}_{\text{sum}}||\text{Att}_{\text{dot}}||\text{Att}_{\text{fc}}
\end{align*}
\]

where + and \cdot are the element-wise addition and multiplication, || is the column-wise concatenation, and \text{fc}(\cdot) is a one-layer fully-connected neural network.

Caption Generator

Given a temporal segment \( S \) in a context \( O \), the caption generator will generate a caption based on \( S \). With the temporal segments generated by the sentence localizer (Sec. 3.2.2), the caption generator first applies soft mask clipping on the contexts, and then uses a context fusion mechanism (Sec. 3.2.3) to fuse the clipped contexts together. The fused contexts are then fed to a caption decoder, which is also a GRU RNN, to generate the corresponding captions.

Soft Mask Clipping. Getting a temporal segment \( S \) from a context, i.e., the clipping operation, is non-differentiable, which makes it difficult to handle in end-to-end training. To this end, we utilize a continuous mask function with regard to the time step \( t \) to perform soft clipping. The mask \( M \) to obtain an \( S \) is defined as follows:

\[
M(t, S) = \sigma(-L(t - c + \frac{L}{2})) - \sigma(-L(t - c - \frac{L}{2}))
\]

where \( \sigma(\cdot) \) is the sigmoid function, and \( L \) is a scaling factor. When \( L \) is large enough, this mask function becomes a step function that performs the exact clipping. We use the normalized weighted sum of the context features (weighted by the mask) as a feature representing \( S \). This operation approximates traditional mean-pooling over clipped frames.

3.2.3 Context Fusion

Because audio and visual representations are from two different modalities, merging them together is a crucial task in a multi-modal setting. We use three different context merging techniques (Fig. 3.3) to fuse the video \( V' \) and audio \( A' \) features obtained after the normalized soft mask clipping operation. We treat \( V' \) and \( A' \) as row vectors.
Multiplicative Mixture Fusion. The multiplicative mixture fusion can make the model automatically focus on information from a more reliable modality and reduce the emphasis on the less reliable one [194]. Given a pair of features \( V' \) and \( A' \), the multiplicative mixture fusion first adds these two contexts and then concatenates the added context with the two original ones. That is, it produces a final context as follows,

\[
C_{\text{final}} = (V' + A')||V'||A' \tag{3.8}
\]

where + and || are the element-wise addition and column-wise concatenation respectively.

Multi-modal Context Fusion. This fusion strategy is similar to Eq. 3.6. But here, we apply the fusion technique on \( A' \) and \( V' \) (segments as opposed to full video context),

\[
C_{\text{final}} = (V' + A')||(V' \cdot A')||fc(V'||A') \tag{3.9}
\]

where || and fc(·) represent column-wise concatenation and one layer of fully connected neural network accordingly.

MUTAN Fusion. MUTAN fusion was first proposed in [28] to solve visual question-answering tasks by fusing visual and linguistic features. We adopt the fusion scheme to fuse \( V' \) and \( A' \). With the idea of Tucker decomposition [311], we first reduce the dimension of \( V' \) and \( A' \),

\[
V'' = \text{tanh}(V' \times W_v) \tag{3.10}
\]
\[
A'' = \text{tanh}(A' \times W_a) \tag{3.11}
\]

where \( W_v \) and \( W_a \) are learnable parameters and \( \text{tanh}(\cdot) \) is the hyperbolic tangent function. Then we produce the final context as follows:

\[
\tilde{C} = ((T_c \times_1 V'') \times_2 A'') \tag{3.12}
\]
\[
C_{\text{final}} = \text{squeeze}(\tilde{C}) \times W_o, \tag{3.13}
\]

where \( T_c \) and \( W_o \) are learnable parameters. \( \times_i, i \in \{1, 2\} \) denotes the mode-i product between a tensor and a matrix, and \( \times \) is the matrix multiplication operation. \( T_c \) models the interactions between the video and the audio modalities, which is a 3-dimension tensor; \( \text{squeeze} \) operator squeezes \( \tilde{C} \) into a row vector.
3.2.4 Training Loss

We follow the training procedure and loss function presented in [63] to train our networks. We employ the idea of cycle consistency [380] to train the sentence localizer and the caption generator and treat the temporal segment regression as a classification problem. The final training loss is formulated as

\[ L = L_c + \lambda_s L_s + \lambda_r L_r \]  

(3.14)

where \( \lambda_s \) and \( \lambda_r \) are tunable hyperparameters. \( L_c \) is the caption reconstruction loss, which is a cross-entropy loss measuring the similarity between two sentences. \( L_s \) is the segment reconstruction loss, which is an L2 loss. It measures the similarity between two temporal segments. \( L_r \) is the temporal segment regression loss, which is also a cross-entropy loss because we regard the temporal segment regression as a classification problem.

3.3 Experiments

In this section, we first describe the dataset used in our experiments, which is an extension of the ActivityNet Captions Dataset [165] (Sec. 3.3.1). Then we present the experimental setup and implementation details (Sec. 3.3.2). Lastly, we discuss the experimental results for both unimodal (i.e., trained using either audio or video modality) and multi-modal approaches (Sec. 3.3.3).

3.3.1 Dataset

ActivityNet Captions dataset [165] is a benchmark for large-scale dense event captioning in videos. The dataset consists of 20,000 videos and each video is annotated with a series of temporally aligned captions. On average, one video corresponds to 3.65 captions. However, besides the captions, the current dataset only provides C3D features [136] for visual frames, no original videos. To obtain the audio tracks for those videos, we needed to find the original videos on YouTube and download the audio via the provided URLs. Around 5,000 videos are unavailable on YouTube now. We are able to find 8026 videos (out of 10009 videos) for training and 3880 videos (out of 4917 videos) for validation. We use those available training/validation videos throughout our experiments.

3.3.2 Experiment Setup and Implementation Details

We follow the experiment protocol in [63] to train and evaluate all the models. We consider the models proposed in [63] as our baselines, i.e., unimodal models that
Table 3.1: **Audio Only Results.** Illustrated are dense captioning results of pre-trained and final models using audio only.

<table>
<thead>
<tr>
<th>Features</th>
<th>M</th>
<th>C</th>
<th>R</th>
<th>B@1</th>
<th>B@2</th>
<th>B@3</th>
<th>B@4</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pretrained model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC</td>
<td>2.70</td>
<td>6.46</td>
<td>6.74</td>
<td>5.52</td>
<td>1.74</td>
<td>0.67</td>
<td>0.21</td>
<td>3.51</td>
</tr>
<tr>
<td>CQT</td>
<td>2.38</td>
<td>5.60</td>
<td>5.72</td>
<td>4.37</td>
<td>1.57</td>
<td>0.46</td>
<td>0.13</td>
<td>2.90</td>
</tr>
<tr>
<td>SoundNet</td>
<td>2.63</td>
<td>5.76</td>
<td>6.99</td>
<td>6.28</td>
<td>1.81</td>
<td>0.38</td>
<td>0.12</td>
<td>3.44</td>
</tr>
<tr>
<td><strong>Final model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC</td>
<td>3.36</td>
<td>9.56</td>
<td>8.51</td>
<td>6.88</td>
<td>2.55</td>
<td>1.23</td>
<td>0.60</td>
<td>4.20</td>
</tr>
<tr>
<td>CQT</td>
<td>3.25</td>
<td>8.97</td>
<td>7.43</td>
<td>6.34</td>
<td>2.69</td>
<td>0.93</td>
<td>0.32</td>
<td>3.63</td>
</tr>
<tr>
<td>SoundNet</td>
<td>3.41</td>
<td>9.21</td>
<td>8.50</td>
<td>7.19</td>
<td>2.15</td>
<td>0.49</td>
<td>0.13</td>
<td>4.22</td>
</tr>
</tbody>
</table>

Table 3.2: **Fusion Strategies.** Testing results for different context fusion strategies for integrating audio and video modalities are illustrated for both pre-trained and final models. We use MFCC audio features and C3D video features for all experiments.

<table>
<thead>
<tr>
<th>Fusion Strategies</th>
<th>M</th>
<th>C</th>
<th>R</th>
<th>B@1</th>
<th>B@2</th>
<th>B@3</th>
<th>B@4</th>
<th>S</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pretrained model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiplicative mixture fusion</td>
<td>3.59</td>
<td>8.12</td>
<td>7.51</td>
<td>7.12</td>
<td>2.74</td>
<td>1.22</td>
<td>0.56</td>
<td>4.58</td>
<td>-</td>
</tr>
<tr>
<td>Multi-modal context fusion</td>
<td>3.55</td>
<td>7.91</td>
<td>7.54</td>
<td>7.24</td>
<td>2.78</td>
<td>1.28</td>
<td>0.62</td>
<td>4.45</td>
<td>-</td>
</tr>
<tr>
<td>MUTAN fusion</td>
<td>3.71</td>
<td>8.20</td>
<td>7.71</td>
<td>7.45</td>
<td>2.92</td>
<td>1.31</td>
<td>0.63</td>
<td>4.78</td>
<td>-</td>
</tr>
<tr>
<td><strong>Final model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiplicative mixture fusion</td>
<td>4.89</td>
<td><strong>13.97</strong></td>
<td>10.39</td>
<td>9.92</td>
<td>4.17</td>
<td>1.85</td>
<td>0.88</td>
<td>5.95</td>
<td>29.87</td>
</tr>
<tr>
<td>Multi-modal context fusion</td>
<td><strong>4.94</strong></td>
<td>13.90</td>
<td>10.37</td>
<td>9.95</td>
<td>4.20</td>
<td>1.86</td>
<td>0.89</td>
<td>5.98</td>
<td>29.91</td>
</tr>
<tr>
<td>MUTAN fusion</td>
<td>4.93</td>
<td>13.79</td>
<td><strong>10.39</strong></td>
<td><strong>10.00</strong></td>
<td><strong>4.20</strong></td>
<td>1.85</td>
<td><strong>0.90</strong></td>
<td><strong>6.01</strong></td>
<td><strong>30.02</strong></td>
</tr>
</tbody>
</table>

only utilize audio or visual features. Due to the difference in the number of videos for training and validation from the original dataset, we run all the experiments from scratch using the PyTorch implementation provided by [63]. The dimensions of the hidden and output layers for all GRU RNNs (audio/video/caption encoders and caption decoders) are set to 512. We also follow [63] to build the word vocabulary (containing 6,000 words) and preprocess the words.

**Training.** Weak supervision means that we do not have ground-truth temporal segments. We first train the caption generator only (pretrained model), and then train the sentence localizer and caption generator together (final model). To train the pre-trained model, we input the entire context sequence (Fake Proposal, $S = (0.5, 1)$). We use the weights of the pre-trained model toinitialize the relevant weights in the final model. For both the pre-trained model and the final model, we

1https://github.com/XgDuan/WSDEC
Table 3.3: Multi-modal Results. Comparison among unimodal and our multi-modal models using MUTAN fusion.

<table>
<thead>
<tr>
<th>Model</th>
<th>M</th>
<th>C</th>
<th>R</th>
<th>B@1</th>
<th>B@2</th>
<th>B@3</th>
<th>B@4</th>
<th>S</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretrained model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unimodal (C3D video feature) [63]</td>
<td>3.66</td>
<td>8.20</td>
<td>7.42</td>
<td>7.06</td>
<td>2.76</td>
<td>1.29</td>
<td>0.62</td>
<td>4.41</td>
<td></td>
</tr>
<tr>
<td>Unimodal (SoundNet audio feature)</td>
<td>2.63</td>
<td>5.76</td>
<td>6.99</td>
<td>6.28</td>
<td>1.81</td>
<td>0.38</td>
<td>0.12</td>
<td>3.44</td>
<td></td>
</tr>
<tr>
<td>Unimodal (MFCC audio feature)</td>
<td>2.70</td>
<td>6.46</td>
<td>6.74</td>
<td>5.52</td>
<td>1.74</td>
<td>0.67</td>
<td>0.21</td>
<td>3.51</td>
<td></td>
</tr>
<tr>
<td>Multi-modal (SoundNet audio + C3D video feature)</td>
<td>3.72</td>
<td>8.02</td>
<td>7.50</td>
<td>7.12</td>
<td>2.74</td>
<td>1.23</td>
<td>0.58</td>
<td>4.46</td>
<td></td>
</tr>
<tr>
<td>Multi-modal (MFCC audio + C3D video feature)</td>
<td>3.71</td>
<td>8.20</td>
<td>7.71</td>
<td>7.45</td>
<td>2.92</td>
<td>1.31</td>
<td>0.63</td>
<td>4.78</td>
<td></td>
</tr>
<tr>
<td>Final model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unimodal (C3D video feature) [63]</td>
<td>4.89</td>
<td>13.81</td>
<td>9.92</td>
<td>9.45</td>
<td>3.97</td>
<td>1.75</td>
<td>0.83</td>
<td>5.83</td>
<td>29.78</td>
</tr>
<tr>
<td>Unimodal (SoundNet audio feature)</td>
<td>3.41</td>
<td>9.21</td>
<td>8.50</td>
<td>7.19</td>
<td>2.15</td>
<td>0.49</td>
<td>0.13</td>
<td>4.22</td>
<td>25.57</td>
</tr>
<tr>
<td>Unimodal (MFCC audio feature)</td>
<td>3.36</td>
<td>9.56</td>
<td>8.51</td>
<td>6.68</td>
<td>2.55</td>
<td>1.23</td>
<td>0.60</td>
<td>4.20</td>
<td>27.16</td>
</tr>
<tr>
<td>Multi-modal (SoundNet audio + C3D video feature)</td>
<td>5.03</td>
<td>14.27</td>
<td>10.35</td>
<td>9.75</td>
<td>4.19</td>
<td>1.92</td>
<td>0.94</td>
<td>6.04</td>
<td>29.96</td>
</tr>
<tr>
<td>Multi-modal (MFCC audio + C3D video feature)</td>
<td>4.93</td>
<td>13.79</td>
<td>10.39</td>
<td>10.00</td>
<td>4.20</td>
<td>1.85</td>
<td>0.90</td>
<td>6.01</td>
<td>30.02</td>
</tr>
</tbody>
</table>

Table 3.4: Results with ground-truth temporal segments.

train them in both unimodal and multi-modal settings. To train unimodal models, we use initial learning rates of 0.0001 and 0.01 for audio and video respectively with a cross-entropy loss. While training our multi-modal models, we set the initial learning rates to 0.0001 for the network parts that have been initialized with the pre-trained weights, and 0.01 for other network components. \(\lambda_s\) and \(\lambda_r\) in Eq. 3.14 are both set to 0.1. We train the networks using stochastic gradient descent with a momentum factor of 0.8.

Testing. To test the pre-trained models, we select one random ground truth description as well as a random temporal segment instead of an entire video, unlike training. For the final models, following [63], we start from 15 randomly guessed temporal segments, and apply one round of fixed-point iteration and the IoU filtering mechanism to obtain a set of filtered segments. Caption generators are applied to the filtered segments together with context features to produce the dense event captions.

Evaluation metrics. We measure the performance of captioning results using traditional evaluation metrics: METEOR (M) [22], CIDEr (C) [318], Rouge-L (R) [190], Spice (S) [8] and Blue@N (B@N) [235]. For score computations, we use official scripts provided by [165]². Where appropriate, we use mean Intersection over Union (mIoU) to measure segment localization performance.

²https://github.com/ranjaykrishna/densevid_eval
Model | M   | C   | R   | B@1  | B@2  | B@3  | B@4  | S  
---|------|------|------|------|------|------|------|---
Unimodal (C3D) [63] | 4.58 | 10.45 | 9.27 | 8.7  | 3.39 | 1.50 | 0.69 | -  
Multi-modal (SoundNet + C3D) | 4.70 | 10.32 | 9.40 | 8.95 | 3.40 | 1.53 | 0.73 | 5.51  
Multi-modal (MFCC + C3D) | 4.78 | **10.53** | **9.60** | **9.23** | **3.62** | **1.69** | **0.82** | **5.56**  

Table 3.5: Pretrained model results on the full dataset.

### 3.3.3 Experiment Results

Since audio features can be represented in a variety of ways [17, 277, 319], finding the best representation is challenging. We conduct experiments on both pre-trained models and final models using different audio representations, *i.e.*, MFCC [138], CQT [35], and SoundNet [17], which are described in Sec. 3.2.1. Table 3.1 shows the experiment results of pre-trained models and final models using only audio features. We can see that both MFCC and soundNet can generate comparable results.

As discussed in Sec. 3.2.3, in the multi-modal setting, choosing a good fusion strategy to combine both audio and video features is another crucial point. Table 3.2 shows the comparison of different context fusion techniques using MFCC audio representations and C3D visual features (Sec. 3.2.1) for both pre-trained models and final models. Among all fusion techniques, we find that MUTAN fusion is the most appropriate one for our weakly supervised multi-modal dense event captioning task. Therefore, we decided to use the MUTAN fusion technique for our multi-modal models when compared to unimodal models. Tab. 3.3 shows the testing results for comparison among unimodal and multi-modal approaches. We can see that our multi-modal approach (both MFCC and SoundNet audio with C3D video features) outperforms state-of-the-art unimodal method [63] in most evaluation metrics. Specifically, the Bleu@3 and Bleu@4 scores lead to 9% and 13% improvement respectively. Comparing unimodal approaches, we are surprised to find that only using audio features achieves competitive performance. We trained our caption generator with GT segments to remove the effect of localization. The results are shown in Table 3.4. We also conduct experiment on the pretrain caption generator using the full dataset where for some videos, audio data is not available (treated as missing data). We consider zero feature vectors for missing audio. The results are shown in Table 3.5. In addition, we randomly selected 15 validation videos and invited 20 people to conduct a human evaluation for comparing our multi-modal model to the visual-only one. The forced choice preference rate for our multi-modal model is 60.67%.

Figure 3.4 demonstrates some qualitative results for both pre-trained models
Figure 3.4: Qualitative Results. Both pretrained and final model results are illustrated of two videos. Captions are from (a) ground-truth; (b) pretrained model trained only using visual features; (c) multi-modal pre-trained model; (d) the final model trained with video features only; (e) our multi-modal final model for dense event captioning in videos.

and final models. It displays the ground-truth captions along with the ones generated by unimodal models and our multi-modal models. The arrow segments indicate the ground truth or detected temporal event segments. We utilize C3D visual features along with audio features. We can see that our multi-modal approaches outperform unimodal ones, both on caption quality and temporal segment accuracy.

Similar to [63], we are suffering from two limitations. One is that sometimes our multi-modal model can not detect the beginning of an event correctly. The other is that most of the time our final model only generates around 2 event captions, which means that the multi-modal approach is still not good enough to detect all the events in the weakly supervised setting. Overcoming these two limitations is the focus of our future work.
3.4 Conclusion

Audio is a less explored modality in the computer vision community. In this work, we propose a multi-modal approach for dense event captioning in a weakly supervised setting. We incorporate both audio features with visual ones to generate dense event captions for given videos. We discuss and compare different feature representation methods and context fusion strategies. Extensive experiments illustrate that audio features can play a vital role, and combining both audio and visual modalities can achieve performance better than the state-of-the-art unimodal visual model.
Chapter 4

Weakly-supervised Audio-visual Sound Source Detection and Separation

Multi-modal, visual, and auditory, perception is an important aspect of research. Human brain has a remarkable ability to isolate specific conversations from a noisy environment, as noted by Cherry through the "cocktail party effect" [46]. At the same time, we can recognize objects and segment regions corresponding to those objects using our visual and auditory systems. We can also imagine how a particular, visually depicted object, may sound. Each object has unique physical properties, some of which can be visually observed, which leads it to generate a unique sound modulated by interactions with other objects and the environment. Therefore, working jointly with auditory and visual cues can be very useful for the recognition of objects, localization of object regions, and separation of sounds they make in an unsupervised way. Separating sounds of each object from a video has a wide range of applications including audio denoising, hearing aids, automated transcription of speech and music, instrument equalization, audio event remixing, and dialog following.

Recent methods for audio-visual source separation [84, 368, 369] utilize "mix-and-separate" approach to train neural network architectures using self-supervision. The paradigm is simple, given a video, mix the audio track by combining the audio channel with one from another video, and train the network to recover the original audio back, conditioned on the visual encoding of corresponding video content. This paradigm effectively synthesizes the "cocktail party effect" by mixing clean sound(s) with others not present in the scene. While effective in training models for a variety of tasks, such as sound source separation [84, 369] and on-/off-screen audio identification [231], this approach implicitly assumes that videos contain single-source sounds and attempts to correlate regions of the video with spectrograms [369]. Co-separation approach recently introduced by Gao et al. [84] addresses single-source limitation, but relies on object detectors trained with an external dataset (Open Images [162]) annotated with bounding boxes for potential
audible objects. However, object annotations are generally difficult and expensive to collect. In addition, while audio classes and corresponding spectrogram segmentations, that correspond to detected regions, are “discovered” during training, the model has no capacity to refine object detectors themselves to be optimal for sound-source separation task; e.g., an entire object region is implicitly assumed to produce the sound.

Inspired by prior work, we aim to address the aforementioned limitations. Specifically, we propose a weakly-supervised audio-visual detection and separation method. Our approach, similar to [84], does not assume single-source video; but, unlike [84], also does not rely on externally trained object detection modules or object-level annotations of any kind. Instead, we leverage weak video-level labels to jointly learn visual and auditory segments that depend on one another. Our architecture has two paths: (1) a video frame semantic segmentation path designed to segment a frame into a set of regions using an attention mechanism that generates per-object-class attention map trained using weak frame-level classification objective; and (2) a spectrogram mask prediction path which takes both mixed spectrogram and pooled object-class image features and outputs a dense spectrogram mask with an objective to mask out the mixed-in sound. The spectrogram mask prediction branch is implemented using attention U-Net architecture [267], similar to [84, 369]. However, importantly, unlike prior methods, we train U-Net to produce a set of base masks from which a final mask is constructed using a set of sparse coefficients predicted from multi-modal audio-visual features. This architecture design takes inspiration from [33]. We find that such bi-linear decomposition is very useful in practice, allowing spectrograms to collaborate in learning a set of auditory sound bases while relying on coefficient predictor to figure out how those should mix for a specific object type. Finally, despite having weaker supervision (no object annotations), compared to [84], we illustrate superior performance on the benchmark MUSIC and cross-dataset performance on AudioSet datasets.

Contributions. Our main contribution is an audio-visual co-segmentation approach for sound source separation, where the network learns both what individual objects look like and sound like, from videos labeled with only, one or more, object labels. This formulation and architecture have a number of appealing properties. Mainly, it does not assume single sound source input data, can be learned in an end-to-end manner, and requires no additional supervision or bounding box proposals. On the technical side, we introduce weakly-supervised object segmentation in the context of sound separation. We also formulate spectrogram mask prediction using a set of learned mask bases which are combined using sparse coefficients conditioned on multi-modal (visual object and auditory) features. Extensive experiments on the MUSIC dataset [369] show that our proposed approach outperforms
state-of-the-art methods on visually guided sound source separation and sound denoising.

4.1 Related Work

**Audio-only sound source separation.** Sound source separation is a challenging problem in speech processing and was first illustrated by the "cocktail party effect" [46]. Sound separation has a wide range of applications, including music/vocal separation [135, 185, 252–254], speech separation and enhancement [77, 147, 183, 300]. Classical approaches for the task include Independent Component Analysis (ICA) [129], sparse decomposition [386], Computational Auditory Scene Analysis (CASA) [66], probabilistic latent variable models [117, 282], local Gaussian modeling [64, 75], kernel additive modeling [199] and Non-negative Matrix Factorization (NMF) [173, 198]. Among traditional approaches, NMF is still widely used for unsupervised source separation [96, 130, 134, 291, 321], but requires extra supervision to get good results. Recently, deep learning-based approaches [111, 127, 294] have gained popularity and shown significant improvement in separation performance over earlier methods. Most of recent methods [111, 128] use "Mix-and-Separate" framework to train the network by artificially mixing multiple audio streams first and learning to separate each audio from the mixture. We also use mix-and-separate idea, but following [84, 368, 369] use visual features to guide audio separation.

**Audio-visual source separation.** Multi-modal learning has recently become a popular topic in the computer vision community. The auditory signal is used to supervise the vision model during training in [233]. Similarly, in [17] visual features are used to guide sound models. Audio and vision jointly are used to train the model in [13, 158]. The same line of research also includes generating sounds from silent video [232, 374]. Following [369], we also use audio-visual features to perform the separation. Unlike [84] we do not use any pre-trained object detector and propose an end-to-end approach to detect, localize, and separate sound sources. Early audio-visual works also explore the relationship between motion and sound. Fisher *et al.* [74] used maximal mutual information, while others focused on canonical correlation methods [133, 148]. Unlike [23, 126, 368] who focus on motion and onset events, our aim is to detect and segment sound-making objects to supervise sound separation tasks.

**Weakly supervised visual learning.** Given a video, our approach is able to detect which audio signals correspond to which objects and localize those objects within the video frames. We consider this in the context of weakly-supervised
Chapter 4. Weakly-supervised Audio-visual Sound Source Detection and Separation

4.1 Weakly-supervised Sound Source Detection and Segmentation

Semantic segmentation, which requires no proposals or any instance-level supervision. Instead, only video-level labels are required. Earlier approaches solve weakly-supervised detection and segmentation using Multiple Instance Learning (MIL) [239, 242] and Expectation-Maximization (EM) procedure [234]. More recently, pseudo-annotation generation [122, 283, 333] has gained popularity. Feedback CNN architectures can use CAM [371] or Grad-CAM [278] to identify discriminative object regions directly, or as a form of pseudo-annotations for strong fully-supervised methods like Faster R-CNN [263] or Mask R-CNN [106], leading to a two-stage approach. Zhang et al. [365] propose an end-to-end pseudo-annotation generation pipeline by introducing a decoupled spatial neural attention network to localize discriminative parts and estimate object regions simultaneously. Motivated by [365], in this work, we generate pseudo-annotations for weakly supervised detection and segmentation; and use these to visually guide the network to separate sound in an end-to-end fashion instead of using pre-trained object detectors [84].

4.2 Approach

We introduce a method to visually guide sound separation using segmented object regions predicted to make sounds. In this section we first formalize our audio-visual sound source separation and detection task (4.2.1) and then focus on the proposed deep network architecture for solving it (4.2.2).

4.2.1 Problem Formulation

Given an unlabeled video \( V \) with accompanying audio \( A(t) \) containing a set of \( N \) objects denoted as \( V = O_1, O_2, \ldots, O_N \). Objects within the video can be treated as sound sources and \( A(t) = \sum_{n=1}^{N} A_n(t) \) where \( A_n(t) \) is the discrete time signal for each object. In this work, our goal is to detect and separate sound \( A_n(t) \) of each object \( O_N \) by using object-level audio-visual supervision.

"Mix-and-Separate" framework [67, 83, 84, 111, 128, 355, 369] is a well-known approach for the task of sound source separation. The idea is to generate an artificially complex auditory signal by mixing multiple individual audio signals and learning to separate each sound of interest from the composition (see Figure 4.1 for illustration). We follow a similar strategy for our training.

Given two input videos \( V_1 \) and \( V_2 \) with accompanying audio \( A_1(t) \) and \( A_2(t) \), respectively, we detect and segment objects, that make sound, from each video using a weakly supervised segmentation network. Then we generate a complex mixed auditory signal \( A_m(t) = A_1(t) + A_2(t) \) by mixing two audio signals \( A_1(t) \)
and $A_2(t)$. Using a short-time Fourier transform (STFT) [91] with $F$-frequency bins, we transformed the mixed signal $A_m(t)$ into a magnitude spectrogram $A^M \in \mathbb{R}_+^{F \times N}$. $A^M$ represents the change of frequency and phase over time in the mixed auditory signals. Suppose $V_1$ contains two objects $O_1'$ and $O_1''$ and corresponding audios $A_1(t)'$ and $A_1(t)''$ accordingly. Similarly, $V_2$ contains one object $O_2$ with accompanying audio $A_2(t)$. Now our goal is to separate sounds $A_1(t)'$, $A_1(t)''$ and $A_2(t)$ of each detected object $O_1'$, $O_1''$ and $O_2$ by predicting a spectrogram mask $\mu_n$ with the supervision of visual cues. To train the network one can use either ratio or binary mask and obtain object level magnitude spectrogram by $A_n = A^M \times \mu_n$. Finally, one can apply Inverse Short-Time Fourier transform (ISTFT) [91] to reconstruct object-level wave-form sounds.

### 4.2.2 Weakly-supervised Audio-Visual Architecture

We propose a weakly-supervised audio-visual detection and separation architecture illustrated in Figure 4.2. Our architecture has two paths: (1) a video frame semantic segmentation path designed to detect objects that have potential to make sounds and segment them out in the frame, using an attention mechanism that generates per-object-class attention map, trained using weak frame-level classification objective (top block in yellow in Figure 4.2); and (2) a spectrogram mask prediction path which takes both mixed audio and pooled object-class image features and outputs a dense mask with an objective to mask out the mixed-in sound (bottom block in Figure 4.2).

We propose an end-to-end approach, unlike [84], to detect and segment objects from the input video frame. The input to our video frame segmenter is an RGB
Figure 4.2: **Weakly-supervised Audio-Visual Architecture.** ResNet-18 followed by a $3 \times 3$ convolution layer is used to extract visual feature ($V_f$) from input video frames. These features are fed to the segmentation network to detect the sound sources. Depending on the classification scores from the segmentation network, we generate soft semantic segmentations by producing a class-specific attention map ($X_m$). We use this attention map to pool features from respective image regions of $V_f$ generating $V_{fm}$. The resultant feature is concatenated with the bottleneck features of attention u-net to generate an audio-visual feature ($W_{AV}$). $W_{AV}$ is passed to the mask coefficient generator to generate $k$ mask coefficients ($M$). At the same time, attention U-Net generates $k$ audio channels ($P$) and combined linearly ($\sigma(PM^T)$) to predict the final audio spectrogram mask guided by visual feature.

The spectrogram mask prediction path is trained to generate a (binary or real-valued) mask that masks out the mixed-in sound. Prior approaches decode the multi-modal encoding of the mixed-audio and visual representation of attended frame [369], or an object region in the frame [84], into a mask directly. Instead, we utilize an attention U-Net architecture to first dynamically generate auditory mask bases from the mixed spectrogram itself. We then generate coefficients for these bases conditioned on the multi-modal features. The final mask is constructed as a coefficient-weighted combination of predicted bases. This decomposition allows shared learning of bases, and focuses visual conditioning on a few coefficients;
Chapter 4. Weakly-supervised Audio-visual Sound Source Detection and Separation

this, we find, significantly improves the performance.

**Video frame semantic segmentation.** We use ResNet-18 [107] as a backbone network followed by a $3 \times 3$ convolution to extract $H \times W$ spatial visual features $V_f \in \mathbb{R}^{1024 \times H \times W}$ from the input video frame. These features are fed to the segmentation network to detect and segment objects. Following [365], our object detection network uses decoupled spatial neural attention to detect and localize salient object regions simultaneously. The segmentation network contains two branches: (1) Expansive attention detector which identifies object regions and generates expansive attention map $A_E \in \mathbb{R}^{C \times H \times W}$; and (2) Discriminative attention detector that predicts the discriminative parts and generates discriminative attention map $A_D \in \mathbb{R}^{C \times H \times W}$. Expansive attention detector consists of a drop-out layer, $1 \times 1$ convolution layer, another drop-out layer, a non-linear activation layer (Eq. 4.1) and a spatial-normalization step (Eq. 4.2). Each element in $A_E$ is defined as follows:

$$\alpha^c_{(i,j)} = F(W^T_c V_f(:, i, j) + b^c), \quad (4.1)$$

$$\alpha^c_{(i,j)} = \frac{\alpha^c_{(i,j)}}{\sum_i \sum_j \alpha^c_{(i,j)}}, \quad (4.2)$$

where $c \in C$ and $F(\cdot)$ denote channel/class and non-linear activation respectively. Discriminative attention detector contains a $1 \times 1$ convolution layer and directly outputs a class-specific object attention map $A_D$. We combine both attentions and generate final attention maps as follows: $X_m = A_E \odot A_D$, where $\odot$ is the element-wise multiplication. Each depth channel of $X_m$ is passed through a spatial average pooling layer to generate a classification score for the corresponding class; this results in $S \in \mathbb{R}^{|C|}$ class scores. Then we apply a multi-label classification loss ($c$-loss) denoted as follows:

$$\mathcal{L}_{c-loss} = - \sum_c y_c \log \frac{1}{1 + e^{-S_c}} + (1 - y_c) \log \frac{e^{-S_c}}{1 + e^{-S_c}}, \quad (4.3)$$

where $y_c$ denotes binary GT label for corresponding $c$-th class and $|C|$ is the number of object classes.

Note that $X_m$ can be interpreted as soft semantic segmentation (actual segmentation can be obtained by thresholding $X^c_m$), with each channel corresponding to a specific object type. We can detect which objects are present, at test time, in a given video frame, by thresholding the classification scores $S$.

**Attention U-Net for audio processing.** Motivated by [369], in this work, we use a time-frequency representation of sound. Therefore, first, we apply a Short-Time
Fourier Transform (STFT) on the input mixture sound to generate the corresponding spectrogram. Then the magnitude of the spectrogram is transformed into a log-frequency scale and used for further processing. Following [227], we use attention U-Net to extract audio features from the log magnitude of the spectrogram. Attention U-Net uses attention gate (AG) to highlight discriminative features while passing through the skip connection. We use 7 convolutions (or down-convolutions) and 7 de-convolutions (or up-convolution) with skip connections in between for attention U-Net. The size of the input spectrogram is $1 \times 256 \times 256$ and the final output of attention U-Net is audio mask bases ($\mathbf{P} \in \mathbb{R}^{k \times 256 \times 256}$) with $k$ channels/bases. In this work, we use 32 as the value of $k$.

**Mask Coefficient Generator.** Following [33], the goal of mask coefficient generator is to predict $k$ mask coefficients: $\mathbf{M} \in \mathbb{R}^k$. In this work, we use audio-visual feature to generate mask coefficients. Based on classification scores, $\mathbf{S}_c$, that is above a certain threshold, $\tau$, from the segmentation network, we select corresponding class-specific attention channel(s) of $\mathbf{X}_m$ and apply weighted pooling on the visual feature $\mathbf{V}_f$ to generate attended visual feature for a corresponding object – $\mathbf{V}_{fm}$. The attended visual feature is concatenated with bottleneck U-Net feature, $\mathbf{A}_f$, to produce audio-visual feature vector $\mathbf{W}_{AV}$. $\mathbf{W}_{AV}$ is fed to the mask coefficient generator to predict $k$ mask coefficient ($\mathbf{M}$). The mask coefficient generator consists of a series of convolution layers with non-linear activations and batch-normalization. In this work, we use ReLU as non-linear activation function. We predict final magnitude of spectrogram, $\mu_A$, by linearly combining $k$ audio mask bases from $\mathbf{P}$ with the mask coefficient $\mathbf{M}$ as follows:

$$\mu_A = \sigma(\mathbf{PM}^T),$$

(4.4)

The predicted magnitude of spectrogram $\mu_A$ is combined with the phase of the input spectrogram. Then we use the inverse STFT to get a waveform of the prediction. Our ultimate goal is to learn spectrogram masks of two types: binary or ratio. Following [369], in case of binary mask we use per-pixel sigmoid cross-entropy loss (i.e., BCE Loss, $\mathcal{L}_{BCE}$, to train the network). Similarly, per-pixel $\mathcal{L}_1$ loss [367] is used to train the network when we use ratio mask.

### 4.3 Experiments

#### 4.3.1 Datasets

**MUSIC dataset.** We evaluate our method using MUSIC (Multimodal Sources of Instrument Combinations) dataset [369]. The dataset contains 685 untrimmed
videos of musical solos and duets which includes 11 instrument categories: accor-
dion, acoustic guitar, cello, clarinet, erhu, flute, saxophone, trumpet, tuba, violin,
and xylophone. MUSIC dataset provides YouTube videos by keyword query. We
use these keywords to download videos. We find that 31 videos are now missing
from YouTube. The train/val/test split of the MUSIC dataset is unavailable. There-
fore, we follow the train/val/test split of [84] where the first/second video in each
category is considered as the validation/test data and the rest is used for training
data.

**AudioSet-SingleSource.** This is a small dataset, assembled in [82], which we only
use for evaluation. The dataset contains single object sounds with corresponding
videos. The dataset consists of 15 musical instruments plus additional sounds pro-
duced by animals and vehicles. For our cross-dataset experiment, we randomly
select 11 out of 15 musical instruments for evaluation. Note, the number of the
instruments are *unseen* by the model – not in the MUSIC dataset that we use for
training.

### 4.3.2 Pre-processing and implementation details

Following [369], we use several pre-processing steps on the MUSIC dataset before
training the model. To reduce the computational cost, we sub-sampled the audio
signals to 11kHz, so that most of the important frequencies of instruments will pre-
serve by degrading slightly the overall audio quality. We sample approximately 6
seconds of audio by random cropping from each untrimmed video. A Hann win-
dow size of 1022 and a hop length of 256 is used to compute STFT and generate
a $512 \times 256$ Time-Frequency audio spectrogram which is further re-sampled on a
log-frequency scale to obtain a $256 \times 256$ Time-Frequency representation. This
representation is used as input to the attention U-Net. We obtain an output pre-
predicted mask and apply an inverse sampling step to convert the mask back to linear
frequency scale of size $512 \times 256$ followed by an inverse STFT to recover wave-
form signal.

We use Pytorch to implement our network. Following [84], we randomly sam-
ple 1-frame to train the model. To process the input video frame, we use ResNet-
18 [107] with two modifications: (1) remove last average pooling layer and $fc$
layer, and (2) add a $3 \times 3$ convolution layer with 1024 output channels. We follow
the experimental protocol of [369] and randomly sample 2 videos from MUSIC
dataset to generate mixture of audio. We use SGD optimizer with momentum 0.9
and learning rate 0.001 to train our network. For ResNet-18, we use learning rate
0.0001 since it is pre-trained on ImageNet. We train the network for 100 epochs
and save the best model based on the validation error.
4.3.3 Experimental results for source separation and detection

Evaluation Metrics. To measure performance we use three widely used metrics for sound separation: Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifact Ratio (SAR). All the results are reported using widely used mir eval library [250].

Baselines. We use the following baselines to compare our quantitative results.

- **NMF-MFCC [291].** An audio-only method that uses Non-negative Matrix Factorization to separate sources using Mel frequency cepstrum coefficients; we report results from [84].

- **AV-Mix-and-Separate.** A simple baseline reported in [84] following the “mix-and-separate” framework to do video-level separation using multi-label hinge loss.

- **Sound-of-Pixels [368].** We use publicly available code\(^1\) to train a 1-frame-based model with train/val/test split of [84] and report the results for fair comparison. We keep default settings for other hyper-parameters.

- **CO-SEPARATION [84].** An object-level audio-visual source separation framework that uses detected objects to separate sound. We can directly compare to their results because we use their train/val/test split to train 1-frame-based model.

**Visually guided sound source separation.** Table 4.1 shows quantitative evaluation of experimental results on MUSIC dataset, using both binary and ratio masks.

\(^1\)https://github.com/hangzhaomit/Sound-of-Pixels
We also include sound separation results with and without weakly-supervised segmentation network, as an ablation, to show the importance of that module in our architecture. We note that additionally removing the mask-coefficient component effectively reduces our model to the Sound-of-Pixels [369] baseline – the reason we do not include this variant. We note that improvements due to our decomposable construction of the mask are very significant (7.26 vs. 9.25 in SDR using binary mask). The improvements due to weakly-supervised detection and segmentation is slightly more modest (8.40 vs. 9.14 in SDR using ratio mask) but are still substantial. Consistent with [84], we find SDR and SIR metric to be most informative.

Figure 4.3 shows corresponding qualitative results. The first and second rows illustrate randomly sampled video mixture pairs and corresponding spectrograms of the mixed sound. The third and fourth rows show ground truth and predicted masks respectively; the fifth and sixth rows show ground truth and predicted separated spectrograms. Finally seventh row illustrated predicted spectrogram generated by running pre-train model from [84]\(^2\). One can clearly see that our method outperforms the state-of-the-art [84] in both quality and sharpness of resulting spectrograms.

**Sound object detection and segmentation.** Our object detection and segmentation utilizes a weakly-supervised network. Importantly, in addition to weakly-supervised loss, audio separation pathway, that depends on the resulting segmentations, provides additional regularization. We measure accuracy of our object detection network by computing multi-class classification accuracy on the MUSIC test set, as reported in Table 4.2 as a function of the threshold \(\tau\). Results illustrate that we can achieve high accuracy of up to 93.69\% and that regularization with ratio mask variant of the audio network is consistently better for visual object detection. We further visualize the segmentation localization qualitatively (since dataset does not contain spatial annotations for quantitative analysis) in Figure 4.4.

<table>
<thead>
<tr>
<th>Threshold value ((\tau))</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary mask</td>
<td>80.30</td>
<td>91.41</td>
<td>93.18</td>
<td>92.68</td>
<td>88.89</td>
</tr>
<tr>
<td>Ratio mask</td>
<td>83.08</td>
<td>91.92</td>
<td>93.69</td>
<td>93.18</td>
<td>89.65</td>
</tr>
</tbody>
</table>

Table 4.2: **Multi-label object classification accuracy.** Performance in (%) on the MUSIC test set.

**Cross-dataset experiments.** We also perform cross dataset testing to evaluate the generality of our method. We do so by measuring the performance of our proposed

\(^2\)https://github.com/rhgao/co-separation
Chapter 4. Weakly-supervised Audio-visual Sound Source Detection and Separation

Figure 4.3: Qualitative audio separation results on MUSIC test set. Test samples, our results and comparison with [84] are shown. See text for details and discussion.

model, trained on MUSIC dataset, by applying it on the AudioSet-SingleSource dataset. The results are presented in Table 4.3. Note the nearly $10 \times$ performance increase in SDR as compared to [369].

Audio separation for unseen objects. We also conduct a small experiment to see how the models perform for separating objects/instruments that the model has not seen during training. The results are presented in Table 4.4. Here, the model never seen some instruments (e.g., Banjo, Marimba) during training on MUSIC dataset but evaluated on those instruments from AudioSet-SingleSource dataset.
Figure 4.4: **Attended object map.** Attended map with class labels from our learned weakly supervised segmentation network which focus on objects that make sounds.

<table>
<thead>
<tr>
<th>Methods</th>
<th>SDR (↑)</th>
<th>SIR (↑)</th>
<th>SAR (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sound-of-Pixels [369]</td>
<td>0.72</td>
<td>20.14</td>
<td>16.10</td>
</tr>
<tr>
<td>Ours(Mask coefficient)</td>
<td>5.11</td>
<td>26.57</td>
<td>13.95</td>
</tr>
<tr>
<td>Ours(Mask coefficient + Seg. Net)</td>
<td>7.19</td>
<td>29.98</td>
<td>12.15</td>
</tr>
</tbody>
</table>

Table 4.3: **Cross dataset evaluation of audio separation.** Evaluation of the model trained using the MUSIC dataset on the AudioSet-SingleSource dataset; using ratio mask. SAR is responsible for capturing absence of artifacts, therefore, if can be higher even when separation results are poor.

In this case the model is relying on similarity of novel instruments to those used in training.

<table>
<thead>
<tr>
<th>Instruments</th>
<th>Sound-of-Pixels [369]</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDR (↑)</td>
<td>SIR (↑)</td>
</tr>
<tr>
<td>Banjo/Electric Guitar</td>
<td>0.03</td>
<td>0.08</td>
</tr>
<tr>
<td>Saxophone/Marimba</td>
<td>3.64</td>
<td>5.30</td>
</tr>
<tr>
<td>Cello/Electric Guitar</td>
<td>0.79</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Table 4.4: **Audio separation for unseen objects.** Toy experiment with cross-dataset setting where the model has never seen some instruments during training on MUSIC dataset.
4.4 Conclusion

In this work, we introduce an end-to-end audio-visual co-segmentation network to separate and detect sound sources without requiring additional supervision or bounding box proposal and solve the problem in a weakly supervised manner from large-scale unlabeled videos. Moreover, our mask coefficient generator facilitates separation conditioned on the output from the segmentation network. Both quantitative and qualitative results show the effectiveness of our proposed method compared to the existing state-of-the-art methods for sound source separation. In the future, we will incorporate motion features in addition to visual features to guide the sound source separation and detection task.
Chapter 5

Human-centric Audio-visual Representation Learning

Multi-modal audio-visual learning [377], which explores and leverages the relationship between visual and auditory modalities, has started to emerge as an important sub-field of machine learning and computer vision. Examples of typical tasks include: audio-visual separation and localization, where the goal is to segment sounds produced by individual objects in an audio and/or to localize those objects in a visual scene [79, 84, 256, 369]; and audio-visual correspondence, where the goal is often audio-visual retrieval [118, 297, 358]. Notably, some of the most recent audio-visual methods [79] leverage human pose key points, or landmarks, as an intermediate or contextual representation. This tends to improve the overall performance of sound separation, as pose and motion are important cues for characterizing both the type of instrument being played and, potentially, over time, the rhythm of the individual piece [79]. It can also serve as an intermediate representation when, for example, generating video from acoustic signals [41, 284].

Most of the existing architectures tend to extract features from the necessary modalities using pre-trained backbones (e.g., CNNs applied to video frames [369], object regions [84], and audio spectrograms; and/or graph CNN for human pose [79]) and then construct problem-specific architectures that often utilize simple late fusion for cross-modal integration in decoding (e.g., to produce spectrogram masks [79, 84, 369]). This is contrary to current trends in other multi-modal problem domains, where over the past few years, approaches have largely consolidated around generic multi-modal feature learning architectures that are task agnostic to produce contextualized feature representations and then fine-tune those representations to a variety of tasks (e.g., visual question answering (VQA) or reasoning (VCR)) and datasets. Examples of such architectures include ViLBERT [203], VL-BERT [295], and Unicoder-VL [179], all designed specifically for visual-linguistic tasks.

Audio-visual representation learning has, in comparison, received much less attention. Most prior works [326] assume a single sound source per video and rely on audio-visual alignment objectives. Exceptions include [237], which relies on proposal mechanisms and multiple-instance learning [307] or co-clustering [124]. These approaches tend to integrate multi-modal features extracted using
pre-trained feature extractors (e.g., CNNs) at a somewhat shallow level. The very recent variants [32, 174, 206] leverage transformers for audio-visual representation learning through simple classification [32] and self-supervised [174] or contrastive [206] learning objectives while only illustrating performance on video-level audio-visual action classification. To the best of our knowledge, no audio-visual representation learning approach to date has explored pose as one of the constituent modalities; nor has shown that feature integration and contextualization at a hierarchy of levels, as is the case for BERT-like architectures, can lead to improvements on granular audio-visual tasks such as audio-visual sound source separation.

To address the aforementioned limitations, we formulate a human-centric audio-visual representation learning architecture, inspired by ViLBERT [203] and other transformer-based designs, with an explicit goal of improving the state-of-the-art in audio-visual sound source separation. Our transformer model takes three streams of information: video, audio, and (pose) keypoints and co-attends among those three modalities to arrive at enriched representations that can then be used for the final audio-visual sound separation task. We illustrate that these representations are general and also improve performance on other auxiliary tasks (e.g., forms of cross-modal audio-visual-pose retrieval). From a technical perspective, unlike ViLBERT and others, our model does not rely on global frame-wise features nor an external proposal mechanism. Instead, we leverage a learned attention to form visual tokens, akin to [256], and leverage weakly supervised objectives that avoid single sound-source assumptions for learning. In addition, we introduce spectrogram mask prediction as one of our pre-training tasks to enable the network to better learn task-specific contextualized features.

**Contributions:** Foremost, we introduce a tri-modal ViLBERT-inspired model, which we call TriBERT, that co-attends among visual, pose keypoint, and audio modalities to produce highly contextualized representations. We show that these representations, obtained by optimizing the model with respect to uni-modal (weakly-supervised) classification and sound separation pretraining objectives, produce features that improve audio-visual sound source separation at large and also work well on other downstream tasks. Further, to avoid reliance on the image proposal mechanisms, we formulate tokenization in the image stream in terms of learned attentional pooling, which is learned jointly. This alleviates the need for externally trained detection mechanisms, such as Faster RCNN and variants. We illustrate competitive performance on a number of granular audio-visual tasks both by using the TriBERT model directly, using it as a feature extractor, or by fine-tuning it.
5.1 Related works

Audio-visual Tasks. There exists a close relationship between visual scenes and the sounds that they produce. This relationship has been leveraged to complete various audio-visual tasks. Based on [377]’s survey of audio-visual deep learning, these tasks can be categorized into four subfields, three of which are addressed in this work and described in the following three subsections.

Audio-visual Sound Source Separation and Localization. Sound source separation and the related task of sound source localization have been studied quite extensively. Previous works studying separation, also known as the cocktail party problem [105], leverage multi-modal audio-visual information [67, 77] to help improve performance with respect to their audio-only counterparts [132, 204]. Examples include learning correlations between optical flow and masked frequencies [52, 74], using graphical models [110], detecting salient motion signals that correspond to audio events [176, 244], and extracting pose keypoints to model human movements [79]. A close connection between separation and localization has also been illustrated [244, 269, 368, 369]. For example, [84, 256] both formulate the task as one of auditory and visual co-segmentation, either with pre-trained object regions obtained by the detector [84] or directly from the image [256]. All of these approaches contain highly specialized architectures with custom fusion schemes. We aim to leverage the flexibility of transformer models to create generalized multi-modal representations that improve on audio-visual tasks.

Audio-visual Representation Learning. The goal is typically to learn aligned representations. The quality of these representations has been shown to greatly impact the overall performance of tasks downstream [29]. A common strategy for representation learning is to introduce a proxy task. In the audio-visual space, past works [13, 14, 231] have trained networks by having them watch and listen to a large amount of unlabeled videos containing both positive samples of matching audio and visual pairs and negative samples of mismatched pairs; the proxy task is binary classification of whether or not the audio and visual match each other. Other proxy tasks include determining whether or not an audio-visual pair is time synchronized [159]; and [175] uses a classification task to identify the correct visual clip or audio stream from a set with negative samples. However, these works rely on the assumption that only one main sound source occurs at a time and everything else is background noise. Our model uses a weakly supervised proxy objective to learn representations for multiple sources of sound (two in experiments) occurring simultaneously and also learns to incorporate pose features.
Audio-visual Correspondence Learning. One of the fundamental tasks in correspondence learning related to our work is cross-modality retrieval. Most prior works focus on audio-visual retrieval [119, 220, 298] and propose learning a joint embedding space where both modalities can be mapped to. In this space, semantically related embeddings are also close to each other and thus retrieval can be performed by selecting the closest embedding to the query from the alternate modality. In our work, we demonstrate that enhanced feature representations obtained by our pretrained model capture aligned semantics and lead to much better cross-modal retrieval.

Visiolinguistic Representation Learning. Our model is inspired by the recent successes of visiolinguistic representations. Most such approaches leverage a combination of uni-modal and cross-modal transformer modules to pre-train generic visiolinguistic representations on masked language and/or multi-modal alignment tasks. For example, [203] proposes separate streams for each modality that communicate with each other through co-attention, while [295] uses a single-stream model that takes both visual and linguistic embeddings as input. In our work, we also leverage co-attention modules to learn joint representations between audio, pose, and vision modalities. However, in addition to extending co-attention, we also focus on reformulating image tokenization and demonstrate the ability to learn with weakly-supervised classification objectives as opposed to masked token predictions.

5.2 Approach

We introduce TriBERT, a network that learns a joint representation of three modalities: vision, pose, and audio. We briefly review ViLBERT, the architecture that inspired TriBERT, in Section 5.2.1. We then describe our TriBERT architecture in Section 5.2.2, including pretraining tasks and objectives.

5.2.1 Reviewing Vision-and-Language BERT (ViLBERT)

Motivated by the recent success of the BERT architecture for transfer learning in language modeling, Lu et al. [203] proposed ViLBERT to represent text and visual content jointly. ViLBERT is a two-stream model for image regions and text segments. Each stream is similar to the BERT architecture, containing a series of transformer blocks (TRM) [317]. Given an image $I$ with corresponding regions-of-interest (RoIs) or bounding boxes $v_0, v_1, \ldots v_N$ and an input sentence $S$ with word tokens $w_0, w_1, \ldots w_T$, the final output representations are $h_{v0}, h_{v1}, \ldots, h_{vN}$.
Figure 5.1: **Our TriBERT Architecture.** We train TriBERT on the MUSIC21 dataset under two training tasks: (1) Classification and (2) Sound separation. We introduce an end-to-end segmentation network for visual embeddings which takes consecutive RGB frames as input and outputs detected object features to feed into vision BERT. Following [79], we use graph CNN to generate pose embeddings as input to pose BERT. For audio, we consider the mixed spectrogram of two audio sources and use a VGGish network to generate audio embeddings to feed into audio BERT. We use classification loss to train individual modalities. Due to the lack of annotation for individual objects, we use a weakly supervised classification loss to train TriBERT for the vision and pose streams. Following prior works on sound separation, we utilize attention U-net [227], which takes a mixed audio spectrogram as input and predicts a spectrogram mask guided by audio-visual-pose features.

and $h_{W0}, h_{W1}, ..., h_{WT}$ for the visual and linguistic features, respectively. To exchange information between the two modalities, the authors introduced a co-attentio- nal transformer layer which computes query ($Q$), key ($K$), and value ($V$) pairs like a standard transformer block. The keys and values from each modality are then fed to the multi-headed attention block of the other modality. The attention block in each stream generates attention-pooled features conditioned on the other modality and outputs a multi-modal joint representation which outperforms single-stream models across multiple vision-and-language tasks.
5.2.2 TriBERT Architecture

The architecture of our proposed TriBERT network is illustrated in Figure 5.1. Inspired by the recent success of ViLBERT in the vision-and-language domain, we modify its architecture to a three-stream network for vision, pose, and audio. Similar to ViLBERT [203], we use a bi-directional Transformer encoder [317] as the backbone network. However, TriBERT also introduces integral components that differentiate its architectural design. First, instead of using bounding box visual features generated by a pre-trained object detector [203] or CNN feature columns [37], TriBERT uses a jointly trained weakly supervised visual segmentation network. Our end-to-end segmentation network takes a sequence of consecutive frames to detect and segment objects, and the corresponding features are pooled and fed as tokens to the visual stream. Second, the pose tokens are characterized by per-person keypoints encoded using a Graph CNN, and the audio token is produced by the VGGish Network [112] applied to an audio spectogram. All three types of tokens form the input to TriBERT, which refines them using tri-modal co-attention to arrive at the final multi-modal representations.

Training TriBERT requires the definition of proxy/pretraining tasks and the corresponding losses (see Section 5.2.2). Specifically, while we adopt token masking used in ViLBERT and others, we are unable to define classification targets per token in our visual and pose streams. This is because we only assume per-video labels (e.g., of instruments played) and no access to how those map to attended sounding regions or person instances involved. To address this, we introduce weakly-supervised classification losses for those two streams. Since only one global audio representation is used, this is unnecessary in the audio stream and standard cross-entropy classification can be employed. Finally, motivated by recent works that show that multi-task pretraining is beneficial for ViLBERT [202], we introduce an additional spectrogram mask prediction pretraining task which predicts spectrogram masks for each individual audio source from the input spectrogram (bottom block, Figure 5.1).

Visual Representations. Unlike [203], we consider input video frames instead of detected object/bounding box features as our visual input and propose an end-to-end approach to detect and segment objects from each individual frame. Figure 5.2 illustrates our visual segmentation network which takes in RGB frames as input. To extract global features, we use ResNet50 [107] as the backbone network followed by a $3 \times 3$ convolution to generate $H \times W$ visual spatial features which are then fed into the segmentation network. Following [365], we use a decoupled spatial neural attention structure to detect and localize discriminative objects simultaneously. The attention network has two branches: (1) Expansive attention detector, which aims
Figure 5.2: **Visual Segmentation Network.** We consider ResNet50 [107] followed by a $3 \times 3$ convolution as our backbone network. It outputs a $H \times W$ spatial visual feature ($V_e$) fed into the segmentation network. Class specific attention map ($A_m$) is generated by the segmentation network and used to pool top two detected object features. We consider the features as visual embeddings for TriBERT.

to detect object regions and generate the expansive attention map $S_E \in \mathbb{R}^{C \times H \times W}$ (top branch of Figure 5.2); and (2) **Discriminative attention detector**, which aims to predict discriminative regions and generate the discriminative attention map $S_D \in \mathbb{R}^{C \times H \times W}$ (bottom branch of Figure 5.2). The expansive attention detector contains a drop-out layer followed by a $1 \times 1$ convolution, another drop-out layer, a non-linear activation, and a spatial-normalization layer, defined as follows:

$$\lambda_{c(i,j)} = F(W^c V_e(:,i,j) + b^c), \quad (5.1)$$

$$\alpha_{c(i,j)} = \frac{\lambda_{c(i,j)}}{\sum_i \sum_j \lambda_{c(i,j)}}, \quad (5.2)$$

where $c \in C$ and $F(\cdot)$ denote number of classes and the non-linear activation function, respectively. The final attention map ($A_m$) is generated as: $A_m = S_E \odot S_D$, where $\odot$ denotes element-wise multiplication. A spatial average pooling is applied on $A_m$ to generate a classification score for each corresponding class and pooled-out top two class features from spatial-visual feature ($V_e$). The resultant $3 \times 2 \times 1024$ visual embeddings are used to train our proposed TriBERT architecture, where $3$ corresponds to the number of frames and $2$ to the number of "objects" per frame.

**Keypoint (pose) Representations.** Our goal is to capture human body and finger movement through keypoint representations. Therefore, we extract 26 keypoints for body joints and 21 keypoints for each hand using the AlphaPose toolbox [69]. As a result, we identify the 2D $(x, y)$ coordinates and corresponding confidence
scores of 68 body joints. Following [79], we use Graph CNN to generate semantic context comprising of those joints. Similar to prior work [347] on action recognition, we construct a Spatial-Temporal Graph Convolutional Network $G = \{V, E\}$ where each node $v_i \in \{V\}$ corresponds to the body joint’s keypoint and each edge $e_i \in \{E\}$ the natural connectivity between those keypoints. We use 2D coordinates of the detected body joints with confidence scores as input to each node and construct a spatial-temporal graph by: (1) connecting human body joints within a single frame according to body structure; and (2) connecting each joint with the same joint from the consecutive frames. This way, multiple layers of spatial-temporal graph convolutions are constructed to generate higher-level features for human keypoints. We use publicly available code\(^1\) to re-train their model on our dataset and extract body joint features of size $2 \times 256 \times 68$ before the final classification layer (corresponding to two person instances). We apply a linear layer to transform these to $3 \times 2 \times 1024$ input embeddings for pose BERT where 3 corresponds to the number of visual frames and 2 to maximum number of persons per frame.

**Audio Representations.** Consistent with prior works, we use a time-frequency representation of the input audio. We apply STFT [91] to generate the corresponding spectrogram and then transform the magnitudes of the spectrogram into the log-frequency scale for further processing. The size of the final input audio spectrogram is $1 \times 256 \times 256$ and is used in two ways: (1) as an audio embedding for audio BERT; and (2) as the input audio for attention U-net for the task of sound source separation, which predicts individual audio spectrogram masks (see Figure 5.1). Before passing to audio BERT, we use a VGGish Network [112] to extract global features for input audio embedding.

**Tri-modal Co-attention.** Recent works [27, 203] propose co-attentional transformer layers to generate effective representations of vision conditioned on language and vice versa. In this work, we introduce a tri-modal co-attentional layer, illustrated to the right, by extending ViLBERT’s co-attentional transformer layers [203]. Given intermediate representations for vision, pose and audio, denoted as $H_V(i)$, $H_P(j)$, and $H_A(k)$, respectively, each stream computes individual query ($Q$), key ($K$), and value ($V$) matrices. The keys and values from two modalities are then concatenated together and fed as input to the multi-head attention block of the third modality. As a result, the block generates attention features conditioned on the other two modalities. We keep the rest of the architecture, such as the feed forward layers, residual connections, etc. the same as a standard transformer block,\(^1\)

\(^1\)https://github.com/yysijie/st-gcn
Chapter 5. Human-centric Audio-visual Representation Learning

Figure 5.3: **Tri-modal Co-attention Layer.** The key and value matrices from two modalities are concatenated together and feed as an input to other modality. The final output of Tri-modal co-attention layer is multi-modal feature for each modality conditioned on other two modalities.

which is then used to generate effective multi-modal features.

**Training Tasks**

We pre-train TriBERT jointly on two tasks: *instrument classification* and *sound source separation*. Our proposed architecture has three separate streams and each stream performs an individual classification task. To train our TriBERT model, we use the MUSIC21 dataset [368], which contains 21 instruments.

**Weakly-supervised Visual and Pose Classification.** Our visual segmentation network generates attention features for input video frames. We then apply a spatial pooling, and the resulting feature vector is fed into the visual BERT. We use a special `<SOS>` token at the beginning of the input frame sequence to represent the entire visual input. Following [203], we apply masking to approximately 15% of the input image regions (see Figure 5.1). The output of the visual BERT is a sequence of hidden representations $h_{v0}, h_{v1}, ..., h_{vN}$ conditioned on the pose and audio modalities. We use $h_{v0}$, corresponding to the `<SOS>` token, to perform classification for the detected objects. Similarly, pose BERT generates a sequence of hidden representations $h_{p0}, h_{p1}, ..., h_{pN}$ conditioned on the visual and audio modalities, and we apply classification based on the `<SOS>` token hidden state. Due to the lack of instance annotations, we cannot use region/pose level supervision. Following [31], we use a weakly-supervised approach to perform region selection and
Audio Classification. Since we do not have a sequence of audio embeddings, we artificially create an audio sequence for computational convenience by repeating the VGGish audio feature to generate a sequence of hidden representations $h_{a0}, h_{a1}, \ldots, h_{aN}$ conditioned on the visual and pose modalities. This is done purely for engineering convenience to allow consistent use of tri-modal co-attention across modalities. We then apply audio classification on the feature corresponding to $<$SOS$>$ token.

Multi-modal Sound Source Separation. We consider sound source separation as one of our initial tasks and follow the "Mix-and-Separate" framework [67, 83, 84, 231, 369], a well-known approach to solve this problem. The goal is to mix multiple audio signals to generate an artificially complex auditory representation and then learn to separate individual sounds from the mixture.

Given two input videos $V_1$ and $V_2$ with accompanying audio $A_1(t)$ and $A_2(t)$, we mix $A_1$ and $A_2$ to generate a complex audio signal mixture $A_m(t) = A_1(t) + A_2(t)$. Suppose $V_1$ has two objects $o_1'$ and $o_1''$ with accompanying audio $a_1'$ and $a_1''$ while $V_2$ has one object $o_2'$ with audio $a_2'$. The goal is to separate sounds $a_1'$, $a_1''$, and $a_2'$ from the mixture $A_m(t)$ by predicting spectrogram masks using attention U-net [227], which takes in the mixed spectrogram as input. Attention U-net contains 7 convolutions and 7 de-convolutions with skip connections. The skip connections use attention gates (AG) and the attention U-net outputs the final magnitude of the spectrogram mask (bottom branch in Figure 5.1) guided by audio-visual-pose features. Following [79], we adopt a self-attention based early fusion between the bottle-neck of attention U-net with the fused features (i.e. concatenation of features) corresponding to the $<$SOS$>$ tokens of three BERT streams. We combine the predicted magnitude of the spectrogram mask from attention U-net with the phase of the input spectrogram and then use inverse STFT [91] to get back the wave-form of the prediction.

Training Objective. We consider weakly-supervised classification for the visual and pose modalities. Following [31], we use two data streams from the initial hidden state of each modality. The first stream corresponds to a class score ($\beta_{\text{class}}$) for each individual region to perform recognition. This is achieved by a linear layer followed by a softmax operation (see Eq. 5.3). The second stream computes a probability distribution ($\beta_{\text{det}}$) for performing a proxy detection. This is done by using another linear layer followed by another softmax operation (see Eq. 5.4) as follows:
where \( h^c \in \mathbb{R}^{C \times |R|}, h^d \in \mathbb{R}^{C \times |R|} \) and \( C \) denotes the number of classes. We then aggregate the recognition and detection scores to predict the class of all image regions as follows: \( \beta^R = \beta_{\text{class}}(h^c) \odot \beta_{\text{det}}(h^d) \), where \( \odot \) denotes an element-wise product of the two scoring metrics. Finally we apply \( \text{BCE-loss} \) [55] to train visual and pose BERT. For audio classification, we consider a classification layer to predict audio classes and similarly apply \( \text{BCE-loss} \) to train audio BERT.

For the sound separation task, our goal is to learn separate spectrogram masks for each individual object. Following [369], we use a binary mask which effectively corresponds to hard attention and use per-pixel sigmoid cross entropy loss (\( \text{BCE-loss} \)) to train the network.

**Implementation Details.** We used PyTorch to implement our network. We consider three random consecutive frames with size \( 224 \times 224 \times 3 \) as our input sequence for visual and pose BERT and use pre-trained ResNet50 [107] to extract global visual features for further processing. For the pose stream, we first predict 2D coordinates of body and finger key points of each frame using AlphaPose [69] and then use graph CNN [347] to generate feature vectors for each keypoint. Similar to prior works [79, 369], we sub-sample audio signals to 11KHz to reduce the computational cost and then select approximately 6s of audio by random cropping. To follow the ”Mix-and-Separate” framework [67, 83, 84, 231, 369], we mix audio inputs and generate a time-frequency audio spectrogram using STFT with a Hann window size of 1022 and a hop length of 256. We then transform the spectrogram into the log-frequency scale to obtain the final \( 256 \times 256 \) time-frequency representation. The transformers for visual/pose and audio have a hidden state size of 1024 and 512, respectively, with 8 attention heads. We use the Adam optimizer with an initial learning rate of \( 1e^{-5} \) and batch size of 12 to train the network on 4 GTX 1080 GPUs for \( 6k \) epochs. Training takes approximately 192 hours.

**Runtime Inference**

We use the MUSIC21 dataset [368] to train our network on two pretraining tasks: classification and sound source separation. We can use this network directly for sound separation on MUSIC21. We also fine-tune the pre-trained TriBERT on
the MUSIC dataset [369] with 11 audio classes, which is a sub-set of the MUSIC21 dataset. We follow a fine-tuning strategy where we modify the classification layer from each pre-trained stream and then train our proposed model end-to-end with a learning rate of $1e^{-7}$ for 1500 epochs while keeping the rest of the hyper-parameters the same as the initial task.

5.3 Experiments

Datasets. We consider the MUSIC21 dataset [368], which contains 1365 untrimmed videos of musical solos and duets from 21 instrument classes for the initial training of our TriBERT architecture. For fine-tuning, we use the MUSIC dataset [369], which is a subset of MUSIC21, containing 685 untrimmed videos of musical solos and duets from 11 instrument classes.

5.3.1 Experiments for Sound Separation

Evaluation Metrics. We use three common metrics to quantify the performance of sound separation: Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifact Ratio (SAR). We report all of the results with the widely used mir_eval library [250].

Baselines. The MUSIC21 dataset contains 1365 untrimmed videos, but we found 314 of those to be missing. Moreover, the train/val/test split was unavailable. As a result, for fair comparison, we trained our baselines [79, 369] with the available videos using an 80/20 train/test split. We use publicly available code² to train ”Sound-of-Pixels” [369]. For ”MUSIC-Gesture” [79], we re-implemented the model by extracting pose features using Graph CNN [347]. Our reproduced results are comparable with those reported³. For the MUSIC dataset, we follow the experimental protocol from [256] and consider their reported results as our baselines.

Quantitative and Qualitative Results. Table 5.2 shows the quantitative results for the sound separation pre-training task on the MUSIC21 dataset. Here, we include the performance of our method and baselines when we use only single-source videos (solos) or multi-source (solos+duets) to train all models. Our TriBERT outperforms (10.09 vs 8.08 for single-source in SDR) baseline models in

²https://github.com/hangzhaomit/Sound-of-Pixels

³The reported SIR score in [79] is 15.81, which is close to our reimplementation of their method which achieves a score of 15.27. Our reproduced SDR score is a bit lower, compared to the 10.12 reported in [79]. However, this is perhaps expected given that 23% of the dataset was missing.
### Table 5.1: Sound separation results on the MUSIC test set.

We use TriBERT with pre-trained weights from the MUSIC21 dataset and then fine-tune this model using the MUSIC train set.

<table>
<thead>
<tr>
<th>Methods</th>
<th>SDR (↑)</th>
<th>SIR (↑)</th>
<th>SAR (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMF-MFCC [291]</td>
<td>0.92</td>
<td>5.68</td>
<td>6.84</td>
</tr>
<tr>
<td>AV-Mix-and-Separate [84]</td>
<td>3.23</td>
<td>7.01</td>
<td>9.14</td>
</tr>
<tr>
<td>Sound-of-Pixels [369]</td>
<td>7.26</td>
<td>12.25</td>
<td>11.11</td>
</tr>
<tr>
<td>CO-SEPARATION [84]</td>
<td>7.64</td>
<td>13.8</td>
<td>11.3</td>
</tr>
<tr>
<td>Mask Co-efficient with seg net [256]</td>
<td>9.29</td>
<td>15.09</td>
<td>12.43</td>
</tr>
<tr>
<td>Ours (after fine-tune)</td>
<td>12.34</td>
<td>18.76</td>
<td>14.37</td>
</tr>
</tbody>
</table>

Table 5.1: **Sound separation results on the MUSIC test set.** We use TriBERT with pre-trained weights from the MUSIC21 dataset and then fine-tune this model using the MUSIC train set.

all evaluation metrics. We then fine-tune our model on the MUSIC dataset with a train/val/test split from [84] (see Table 5.1). Our model again outperforms all baselines in all metrics (12.34 vs 9.29 in SDR). Figure 5.4 and 5.5 illustrate the corresponding qualitative results. In both figures, the first and second rows show input mixed video pairs with accompanying audio respectively. The 3rd, 4th and 5th rows present ground-truth mask and predicted spectrogram mask generated by ours method and Music Gesture [79] respectively. Finally, ground truth and predicted spectrogram using our method and Music Gesture are shown in the 6th, 7th and 8th rows accordingly. One can clearly see that predicted mask and spectrogram generated by our method are closer to ground truth.

<table>
<thead>
<tr>
<th>SDR (↑)</th>
<th>SIR (↑)</th>
<th>SAR (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.73</td>
<td>12.11</td>
<td>10.10</td>
</tr>
<tr>
<td>6.72</td>
<td>14.03</td>
<td>9.68</td>
</tr>
<tr>
<td>7.66</td>
<td>14.54</td>
<td>11.06</td>
</tr>
</tbody>
</table>

Table 5.2: **Sound separation results on the MUSIC21 test set.** SDR / SIR / SAR are used to report performance. Separation accuracy is captured by SDR and SIR; SAR only captures the absence of artifacts.

#### 5.3.2 Multi-modal Retrieval

**Retrieval Variants.** In this experiment, we analyze the semantic alignment between the 3 modalities that TriBERT learns to encode. This is done through cross-modal retrieval, where given a single or a pair of modality embeddings, we attempt to identify the matching embedding from the remaining modality. We consider 5 variants: audio → vision, vision → audio, audio → pose, pose → audio, and vi-
Figure 5.4: Qualitative sound separation results on the MUSIC21 test set. Here, we show a comparison between our results and Music-Gesture [79]. See text for details.

Throughout this section, we refer to the embedding we have as the query embedding and the embedding we want to retrieve as the result embedding. We train and evaluate on the MUSIC21 dataset, using the same 80-20 train-test split used to learn TriBERT.

We consider 2 types of embeddings for the 3 modalities. First, we use the transformer-based embeddings, consisting of the concatenations of the hidden rep-
Figure 5.5: **Qualitative sound separation results on the MUSIC21 test set.** Here, we show a comparison between our results and Music-Gesture [79].

representations $h_{v0\ldots v3}$, $h_{p0\ldots p3}$, and $h_{a0\ldots a3}$ for visual, pose, and audio, respectively. Additionally, we establish a baseline by training with the embeddings used as input to the three BERT streams. This baseline can be viewed as an ablation study for the transformer layers.

**Negatives Sampling Details on Cross-Modal Retrieval.** Recall that for the $n$-way multiple choice setting, $n - 1$ choices are negative pairs and only one pair is
Figure 5.6: **Overview of the training scheme for the cross-modal retrieval module in the $n = 4$ case.** Each multiple choice consists of the correct vision+audio fusion embedding along with a pose embedding. The 4 choices: 1 containing the correct pair, 2 containing easy negatives, and 1 containing a hard negative is passed as input into an MLP network. The network outputs a score for each pair, after which a softmax is applied across all 4 scores. We train this network in a siamese fashion (i.e. each MLP block in the figure shares the same weights) using cross entropy loss. In our experiments, we use a 4-layer MLP with Tanh activation.

positive. Accordingly, for $n = 4$, 3 distractors are sampled, each with an incorrect pose embedding, while the 4th choice contains the matching pose embedding for the given vision and audio embeddings. In other words, the fusion embedding consisting of the vision and audio embeddings is kept as the anchor while negatives are sampled from the pose embeddings only. Of the 3 negative pose embeddings, 2 are considered “easy” negatives, sampled randomly from the entire training set, while the last one is a “hard” negative, sampled randomly from a pool of 25 embeddings corresponding to the 25 nearest neighbours of the anchor vision embedding. In the $n = 3$ case, 2 hard negatives and no easy negatives are used, with the same nearest neighbour sampling method based on the anchor embedding. Figure 5.6 shows a diagram of the training scheme for the cross-modal retrieval module.

**Retrieval Training.** Similar to [203], we train using an $n$-way multiple-choice setting. Here, $n$ depends on the variant of the retrieval task, where $n = 4$ for the vision+audio to pose variant and $n = 3$ for the four remaining single-modality
variants. In either case, one positive pair is used and $n - 1$ distractors (hard negatives) are sampled. Further details are provided in the Supplemental Materials. We use an MLP that takes as input a fusion representation of both the query and result embeddings, computed as the element-wise product of the two. The module then outputs a single logit, interpreted as a binary prediction for whether the query and result embeddings are aligned. For the (vision+audio $\rightarrow$ pose) variant, an additional MLP, based on [79], is used to combine the vision and audio embeddings before the final element-wise product with the pose embedding. Additionally, since both the transformer-based and pre-transformer embeddings are not consistent in shape, we also use a linear layer to transform them to a consistent one. This overall retrieval network is trained end-to-end. For each multiple choice, the network computes an alignment score, after which a softmax is applied across all $n$ scores. We train using a cross-entropy loss for 750 epochs with a batch size of 64, using the Adam optimizer with an initial learning rate of $2e^{-5}$.

**Retrieval Results.** Figure 5.7 shows qualitative results for two variants of retrieval. Additionally, Table 5.3 shows quantitative results for the 5 retrieval variants using the transformer-based representation, the baseline pre-transformer representation, and also a model that simply selects randomly from the pool. We see that retrieval using the transformer-based embeddings results in significantly better performance than the pre-transformer ones. This shows that the Tri-modal co-attention modules are an integral component in learning a semantically meaningful relationship between the three modalities.

### Table 5.3: Multi-modal retrieval on the MUSIC21 test set. Top-$k$ accuracy results for 5 retrieval variants. For each variant, retrieval on both the transformer-based (bolded) and pre-transformer (unbolded) embeddings were evaluated. Also shown is the accuracy of a random selection model.

<table>
<thead>
<tr>
<th>Retrieval Variant</th>
<th>Top-1 Accuracy (random = 0.48)</th>
<th>Top-5 Accuracy (random = 2.38)</th>
<th>Top-10 Accuracy (random = 4.76)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio $\rightarrow$ Vision</td>
<td><strong>68.10</strong> 1.43</td>
<td><strong>93.81</strong> 10.48</td>
<td><strong>98.57</strong> 17.62</td>
</tr>
<tr>
<td>Vision $\rightarrow$ Audio</td>
<td><strong>59.52</strong> 4.76</td>
<td><strong>89.52</strong> 18.57</td>
<td><strong>92.38</strong> 31.90</td>
</tr>
<tr>
<td>Audio $\rightarrow$ Pose</td>
<td><strong>63.81</strong> 2.38</td>
<td><strong>86.67</strong> 6.67</td>
<td><strong>94.29</strong> 10.95</td>
</tr>
<tr>
<td>Pose $\rightarrow$ Audio</td>
<td><strong>54.29</strong> 1.90</td>
<td><strong>85.24</strong> 9.05</td>
<td><strong>86.67</strong> 14.29</td>
</tr>
<tr>
<td>Vision + Audio $\rightarrow$ Pose</td>
<td><strong>54.29</strong> 3.33</td>
<td><strong>90.00</strong> 9.05</td>
<td><strong>96.19</strong> 15.24</td>
</tr>
</tbody>
</table>
Figure 5.7: **Qualitative cross-modal retrieval results on the MUSIC21 test set.** The similarities (same instrument class) between results in the top-5 retrieval pool show that our transformer-based representations have learned a semantically meaningful relationship between the modalities.

### 5.4 Conclusion

In this work, we introduce TriBERT, a three-stream model with tri-modal co-attention blocks to generate a generic representation for multiple audio-visual tasks. We pre-train our model on the MUSIC21 dataset and show that our model exceeds state-of-the-art for sound separation. We also find that TriBERT learns more generic and aligned multi-modal representations, exceeding the cross-modal audio-visual-pose retrieval task. In this work, we limit ourselves to two datasets and fundamental audio-visual tasks. In the future, we plan to consider using more datasets and expanding to a broader set of tasks (e.g., generation). The role of positional embeddings should also be explored.
Chapter 6

Visual Memory Conditioned Consistent Story Generation

Figure 6.1: Referential and consistent story visualization. Examples of more natural stories with references for the FlintstonesSV [97] and MUGEN [104] datasets (bottom text) compared to more typical but less natural story text (top). We extend the MUGEN dataset by introducing additional two characters (e.g., Lisa and Jhon) and four backgrounds (e.g., Sand, Grass, Stone, and Dirt).

Multimodal deep learning approaches have pushed the quality and the breadth of conditional generation tasks such as image captioning [140, 208, 213, 320, 330] and text-to-image synthesis [143, 262, 353, 361, 362, 382]. Owing to the technical leaps made in generative models, such as generative adversarial networks (GANs) [90], variational autoencoders (VAEs) [153] and the more recent diffusion models [114], approaches for text-to-image synthesis can now generate images with high visual fidelity representative of the textual descriptions. The captions, however, in such cases, are generally short self-contained sentences representing the high-level semantics of a scene. This is rather restrictive in the real-world applications [57, 186, 187] where fine-grained understanding of object interactions, motion, and background information described by multiple sentences becomes necessary. One such task is that of story generation or visualization – the goal of which is to generate a sequence of illustrative image frames with coherent semantics given a sequence of sentences [186, 210, 211, 359].

Characteristic features of a good visual story are high visual quality over multiple frames; this includes rendering of discernible objects, actors, poses, and realistic interactions of those actors with objects and within the scene. Moreover, for
text-based story generation, it is crucial to maintain consistency between the generated frames and the multi-sentence descriptions. Not only the actor context but also the background of the generated story should be in-line with the description demonstrating effortless transition and adaptation to the changing environments within the story [209].

Recent advances on the task of story generation have made significant advances along these lines, showing high visual fidelity and character consistency for story sentences that are self-contained and unambiguous (explicitly mentioning characters and the setting each time). While impressive, this setup is fundamentally unrealistic. Realistic story text is considerably more complex and referential in nature; requiring the ability to resolve ambiguity and references (or co-references) through reasoning. As shown in Fig. 6.1, while the description corresponding to the first frame has an explicit reference to the character names, the (typical) subsequent frame descriptions, provided by human, contain references such as “she, he, they”. Moreover, while maintaining character consistency, current approaches are limited in preserving, or transitioning through, the background information in agreement with the text (cf. Fig. 6.3) [186, 210, 211].

In natural language processing (NLP) co-reference resolution in text is an important and core task [11, 146, 215]. While it may be possible to apply such methods to story text to first resolve ambiguous references and then generate corresponding images using existing story generation approaches, this is sub-optimal. The reason is that co-reference resolution in the text domain, at best, would only allow to resolve references and maintain consistency across identity of the character. Appearance across frames would still lack consistency and require some form of visual reasoning. As also noted in [280], reference resolution in the visual domain, or visio-lingual domain, is more powerful.

In this work, for the first time (to our knowledge), we study co-reference resolution in story generation. Prior work [210] offers limited performance when faced with text containing references (see Sec. 6.4). We address this by proposing a new autoregressive diffusion-based framework with a visual memory module that implicitly captures the actor and background context across the generated frames. Sentence-conditioned soft attention over the memories enables effective visio-lingual co-reference resolution and learns to maintain scene and actor consistency when needed. Further, given the lack of datasets that contain references and more complex sentence structure, we extend the MUGEN dataset [104] and introduce additional characters, backgrounds, and referencing in multi-sentence storylines.

Contributions. Our contributions are three-fold: (i) First, we introduce a novel autoregressive deep generative framework, Story-LDM, that adopts and extends la-
tent diffusion models for the task of story generation. As part of Story-LDM, we propose a meticulously designed memory-attention mechanism capable of encoding and leveraging contextual relevance between the part of the story-line that has already been generated, and the current frame being generated based on learned semantic similarity of corresponding sentences. Equipped with this, our sequential diffusion model can generate consistent stories by resolving and then capturing temporal character and background context. (ii) Second, to validate our approach for co-reference resolution, and character and background consistency in the visual domain, we extend existing datasets to include more complex scenarios and, importantly, referential text. Specifically, we extend the MUGEN dataset [104] to include multiple characters and diverse backgrounds. We also modify FlintstonesSV [97] and PororoSV [186] dataset to include character references. These enhancements allow us to increase the complexity of the aforementioned datasets by introducing co-references in the sentences of a story. (iii) Finally, to evaluate different approaches for foreground (character) as well as background consistency we propose novel evaluation metrics. Our results on the MUGEN [104], the PororoSV [186], and the FlintstonesSV [97] datasets show that we outperform the prior state-of-the-art on consistency metrics by a large margin.

6.1 Related work

Text-to-image synthesis. Deep generative models, particularly, generative adversarial networks (GANs) [90], variational autoencoders (VAEs) [153] and normalizing flows [30, 61, 62] have been applied to multimodal tasks at the intersection of vision and language. Typical such tasks include image captioning [9, 208, 329] and text-to-image synthesis [207, 262, 310, 353, 361, 382]. Early work on text-conditioned image synthesis built upon the success of GANs [262]. More recent approaches have utilized multi-stage generators [361] and normalizing flow-based priors [207] in the latent space to model the distribution of images given text. Various approaches have found the cross-domain contrastive loss to improve text-to-image generation models [143, 362]. DALL-E [260] and Cogview [60] harness the power of transformers [317] and discrete variational autoencoders (VQ-VAE) [261] yielding very high-quality image samples.

More recent are the advances in diffusion models which have revolutionized the domain of image generation [114]. Diffusion models progressively add noise to the data and learn a reverse diffusion process to reconstruct it. Nichol et al. [223] adapted diffusion models for text-to-image generation and explored CLIP [248] guided generation as well as classifier-free modeling. Standard diffusion models are employed directly in the high-dimensional pixel space and therefore,
cannot directly be used for the more complex task of story generation. Recent work [44, 93, 265] instead use encodings from pre-trained models as input to the diffusion models, thereby reducing the complexity of the task by working in a lower-dimensional space. In this work, we build upon this idea and extend it for sequential story generation.

**Text-to-video synthesis.** One of the challenges of text-to-video synthesis is the smoothness of motion in a video [186]. Early work focused on generating short clips [57, 187]. To effectively learn the motion, various approaches disentangle the motion features from the background information [102, 312, 322]. Wu et al. [335] propose a novel two-dimensional VQ-VAE and sparse attention module for real-
Chapter 6. Visual Memory Conditioned Consistent Story Generation

world text-to-video generation. Singer et al. [286] decompose the temporal U-Net [268] and the attention modules to approximate them in space and time to extend the text-to-image diffusion models to model text-to-video generation. Ligong et al. [100] propose a transformer framework to jointly model various modalities.

Story Generation. Li et al. [186] proposed the initial idea and task of story generation. A two-level StoryGAN framework is applied to ensure image-level consistency between each sentence and image pair, and a global discriminator enforces global consistency between the entire image sequence and the story. Various approaches have proposed improvements to the StoryGAN architecture. Zeng et al. [359] introduce sentence-level alignment and word-based attention to improve relevance. Li et al. [178] further improve the performance with enhanced discriminators and dilated convolutions. In [289] foreground-background information is provided as additional supervision and [209] uses video captioning for semantic alignment between text and frames. Recently, Chen et al. [39] adopted visual planning and character token alignment to improve character consistency.

Story Completion. Recently, another task for text-to-story synthesis referred to as story completion has been proposed [211]. In this task, in addition to sentences, the first frame of the story is provided as input. In effect, story completion is a simplified variant of story generation. StoryDALL-E [211] leverages models pre-trained for text-to-image synthesis to perform story completion. Datasets for this task include CLEVR-SV [186] and Pororo-SV [97] which are derived from the CLEVR dataset [139], and the Flintstones dataset for text-to-video synthesis has also been modified for the task of story visualization [209]. Additionally, to evaluate the generalization performance, the popular DiDeMo dataset for video captioning [10] is adapted for the task in [211].

Reference Resolution. Co-reference resolution is an important and well-researched topic in NLP and focuses on resolving the pronouns and their associated entities. Classic methods in NLP to co-reference resolution employ decision trees [11, 215], maximum-entropy modeling [146], cluster-ranking [255] and classification algorithms [290]. More recent approaches [71, 141, 156] leverage neural network architectures to obtain improved performance with Transformers [58]. Seo et al. [280] proposed visual co-reference resolution for the task of Visual Question-Answering (VQA) dialogs. We take inspiration from [280], but propose a much more sophisticated memory-attention module that allows us to perform visio-lingual co-reference resolution (and visual consistency modeling) for visual story generation.
6.2 Approach

To generate temporally consistent stories based solely on the linguistic storyline, we develop a deep generative approach with autoregressive structure. We build upon the success of diffusion models in modeling the underlying data distribution of images to produce high-quality samples and learn the generative conditional distribution of the visual story based on the textual descriptions. Given that the multi-frame stories involve high-dimensional data input, we employ Latent Diffusion Models [265], such that diffusion models can be applied in a computationally efficient manner. Besides, to ensure temporal consistency and smooth story progression, we propose a novel memory attention mechanism that not only attends to the multimodal representations of the current frame but also takes into account the already generated semantics of the previous frames. This module also allows us to resolve ambiguous references (e.g., he/she, they, etc.) using visual memory and is the core of our technical contribution. We first provide an overview of the Diffusion Models and the Latent Diffusion Models, following which we present our autoregressive latent diffusion model for stories called Story-LDM1.

6.2.1 The Latent Diffusion Model Backbone

Diffusion Models. Diffusion models are a class of generative models that approximate the underlying data distribution \( p(x) \) by denoising a base (Gaussian) distribution in multiple steps using a reverse process of a fixed Markov Chain of length \( T \). To estimate \( p(x) \), the forward diffusion process starts from the input data \( x_0 = x \) and gradually adds noise to obtain a set of noisy samples \( x_1, \ldots, x_T \) such that \( x_T \sim \mathcal{N}(0, 1) \) represents a sample from a Gaussian distribution. Under the Markov assumption, the probability of the forward process modeling the distribution \( q(x_0:T \mid x_0) \) and the reverse diffusion process estimating probability at an earlier time-step are formulated as:

\[
q(x_1:T \mid x_0) := \prod_{i=1}^{T} \mathcal{N}(x_i; \sqrt{1 - \beta_i}x_{i-1}, \beta_i I) \\
p_\theta(x_0:T) = p_\theta(x_T) \prod_{i=1}^{T} p(x_{i-1} \mid x_i).
\]

(6.1)

Here, \( \{\beta_i\}_{i=1}^{T} \) is the variance schedule for each time-step such that \( x_T \) is nearly a Gaussian. The model parameters \( \theta \) are learned with the following objective,

\[
\mathcal{L}_{DM} := \mathbb{E}_{t,x,\epsilon} \left[ ||\epsilon - \epsilon_\theta(x_t, t)||_2^2 \right],
\]

(6.2)

1https://github.com/ubc-vision/Make-A-Story
where $\epsilon \sim \mathcal{N}(0, 1)$ and $\epsilon_{\theta}(x_t, t)$, $t = 1, \ldots, T$ is a sequence of denoising autoencoders with noisy input $x_t$ predicting the noise that was added to the original input $x$.

Despite yielding state-of-the-art results in various image generation tasks, diffusion models directly operating in the high-dimensional pixel are computationally expensive and resource-exhaustive. This limits their application to even higher-dimensional data such as multi-frame stories or video datasets, which is the focus of this work.

**Diffusion Models in the Latent Space.** To broaden the applicability of the diffusion models to very high-dimensional data e.g. high-resolution images, Latent Diffusion Models (LDM) [265] first compress the original image to a lower-dimensional space using perceptual image compression. An auto-encoder approach is employed such that the original spatial structure of the input image is preserved in the latent space. That is, the encoder $E(\cdot)$ maps the input image $I \in \mathbb{R}^{H \times W \times 3}$ to a latent representation $Z \in \mathbb{R}^{h \times w \times c}$, downsampling the image to a lower spatial dimension. Following this, the diffusion model is applied to the latent $Z$, where time-conditioned U-Net $\epsilon_{\theta}(Z_t, t)$ is employed to model the diffusion process. The objective of the diffusion model from Eq. (6.2) becomes,

$$L_{LDM} := \mathbb{E}_{E(z), \epsilon} \left[ \| \epsilon - \epsilon_{\theta}(Z_t, t) \|_2^2 \right],$$

(6.3)

During training, a forward diffusion process is applied to generate $Z$, which are mapped to the original image space using a decoder $D(\cdot)$.

### 6.2.2 Story-Latent Diffusion Models

Given a textual story, characterized by sequence of $M$ sentences $S_{txt} = \{S^0, \ldots, S^M\}$, the goal of story generation is to produce a sequence of corresponding frames $S_{img} = \{I^0, \ldots, I^M\}$ that visualize the story. We note that this is a more difficult problem than one of story continuation [211], where in addition to the textual story $S_{txt}$ approaches have access to a source frame $I^0$ for additional context at inference time. During training, it is assumed that we have access to a paired dataset of $N$ samples $D = \{S_{txt}^{(i)}, S_{img}^{(i)}\}_{i=1}^N$. We extend latent diffusion models to this task, by allowing them to generate multi-frame stories autoregressively, and by introducing rich conditional structure that takes into account the current sentence as well as context from earlier generated frames through a visual memory module. This visual memory allows the model to incorporate character/background consistency and resolve text references when needed, resulting in improved performance.

Given a condition $y$, LDM utilizes a cross-attention layer with key $(K)$, query
Table 6.1: Dataset statistics of the MUGEN, FlintstonesSV and PororoSV.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Ref (avg.)</th>
<th># Chars</th>
<th># Backgrounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUGEN [104]</td>
<td>None</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Extended MUGEN</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>FlintstonesSV [97]</td>
<td>3.58</td>
<td>7</td>
<td>323</td>
</tr>
<tr>
<td>Extended FlintstonesSV</td>
<td>4.61</td>
<td>7</td>
<td>323</td>
</tr>
<tr>
<td>PororoSV [186]</td>
<td>1.01</td>
<td>9</td>
<td>None</td>
</tr>
<tr>
<td>Extended PororoSV</td>
<td>1.16</td>
<td>9</td>
<td>None</td>
</tr>
</tbody>
</table>

Table 6.2: Quantitative results. Experimental results on the FlintstoneSV and the MUGEN datasets.

\[
\text{Attention}(K, Q, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d}} \right) V. \tag{6.4}
\]

Here, \( Q = W_Q \hat{f}(Z) \), \( K = W_K f(y) \) and \( V = W_V f(y) \), and \( W_Q \in \mathbb{R}^{d \times d_q} \), \( W_K \in \mathbb{R}^{d \times d_k} \) and \( W_V \in \mathbb{R}^{d \times d_v} \) are learnable parameters, \( \hat{f}(Z) \) an intermediate flattened feature representation of \( Z \) within the diffusion model and \( f(y) \) the feature representation of the condition \( y \). The objective in Eq. (7.1) for conditional generation becomes,

\[
\mathcal{L}_{\text{LDM}} := \mathbb{E}_{t, E(I), \epsilon} \left[ \| \epsilon - \epsilon_\theta(Z_t, f(y), t) \|_2^2 \right]. \tag{6.5}
\]

Note that the denoising autoencoders \( \epsilon_\theta \) now additionally depend on the condition encoding \( f(y) \).

For a sample \( \{ S_{\text{txt}}^{(i)}, S_{\text{img}}^{(i)} \} \), we first project all the \( M \) input frames of the story, \( I^0, \ldots, I^M \) onto a low-dimensional space using a frame-encoder \( E \), and obtain the encoded frames \( Z^0, \ldots, Z^M \) for a single story\(^2\). To effectively condition on the corresponding frame, we apply this cross-attention layer to the intermediate

\(^2\text{We drop the superscript denoting sample } i \text{ for ease of notation.}\)
Chapter 6. Visual Memory Conditioned Consistent Story Generation

Lisa moves left then right onto a ladder before dismounting the ladder onto the right of the top platform. It then jumps twice to the right onto a box then onto a coin before walking right into a gear being slain in Grass.

She walks to the right and collects a gem and a coin then stutters while walking towards a ladybug.

She walks to the left. Then it jumps to the right, onto a platform, then it collects a coin.

Figure 6.3: Story generation result for MUGEN. Here we can compare our method and LDM [265]. See text for details.

represents of the neural network for each frame $I^m$ and its textual description $S^m$ using Eq. (7.2), where the relevance of the description $S^m$ is weighted by the similarity between the textual representation and the encoded frame representation $Z^m$ (cf. Fig. 6.2a).

For sequential generation, the model in addition to the current state, requires information from all the previous states. To enable this, the diffusion process for any frame representation $Z^m$ is conditioned on the visual representations of the previous frames $Z^0, \ldots, Z^{m-1}$ as well as the sentence descriptions $S^0, \ldots, S^m$. This conditioning is realized through a novel Memory-attention module which forms the basis of our autoregressive approach.

Memory-attention Module. To capture the spatio-temporal interactions across multiple frames and sentences for a story, in our conditional diffusion model, we condition the frame $Z^m$ not only on the corresponding text $S^m$ but also on the previous texts $S^i$, for $i \in \{0, \ldots, m-1\}$. This conditioning is applied throughout the $T$ time-steps of the diffusion process for $Z^m$. The conditional denoising autoencoder thus models the conditional distribution $p(Z^m | Z^{<m}, S^{<m})$. The (conditional) generative process of our Story-LDM approach over the $T$ steps of the diffusion...
Chapter 6. Visual Memory Conditioned Consistent Story Generation

Fred and Barney are having a conversation on the couch in the living room. Fred turns his head with a disdained look. They are sitting in a room. Fred looks angry and sticks his tongue out while he and Barney are talking.

The circus performer on the television talks to his audience.

Fred is on tv in the room wearing a super hero outfit.

Wilma and Fred are standing in the room. Wilma is talking to Fred. Fred is holding a suitcase and clothing.

Fred is standing in a room speaking with an exasperated expression. He is holding a pile of clothes.

Wilma is standing in a room, she blinks twice before speaking. She is standing in a room, talking to someone. Then, she giggles.

Figure 6.4: Story generation results for FlintstoneSV. Our method is able to generate more consistent characters/backgrounds.

The key motivation for this approach is to propagate the semantic (visual and textual) features from the already processed storyline based on the relevance of the current description to the previous frames as well as the previous descriptions. We achieve this by implementing a special attention layer, called memory-attention module. Similar to the cross-attention layer, we utilize the attention mechanism based on the key, query, and value formulation. In this case Eq. (7.2) becomes,

$$Q = W_Q.f(S^m), \quad K = W_K.f(S^{<m}), \quad V = W_V.\hat{f}(Z^{<m}),$$

where $\hat{f}(\cdot)$ is applied to align the dimensions of the values, $V$ with the keys, $K$. In the memory attention module, the relevance of the query $Q$ which depends on the current sentence $S^m$ and the keys $K$ which represent the previous sentences $S^{<m}$ is used to weight the feature representations $Z^{<m}$. The aggregated representation now contains the information relevant for the current frame $Z^m$ from the already
generated storyline (see Fig. 6.2b). That is, our mechanism based on the similarity of the current sentence to the previous sentences in the story, identifies the features in the previous frames which are of importance to the context of the current frame. This may include the recurrence of certain semantics within the story such as characters or backgrounds. In all, this formulation of the diffusion process allows us to maintain temporal consistency as we amplify the visual feature information from the sequence of stories already generated. This allows the model to implicitly capture temporal dependencies in storylines for resolving ambiguities in character and background information.

Given the above conditioning, the objective for the story latent diffusion model for a single frame is formalized as

$$
\mathcal{L}_{story-LDM} = \mathbb{E}_{Z_{t}, \epsilon, t} \left[ \| \epsilon - \epsilon_{gm}(Z_{m}^{<m}, S^{<m}, Z^{<m}, t) \|_{2}^{2} \right], \quad (6.8)
$$

where $\epsilon_{gm}$ are the denoising autoencoders for the frame $m$.

Having formalized the diffusion process for single frame generation, the generative process for the entire storyline using Eq. (6.6) for autoregressive conditional frame generation, is given by

$$
p(Z_{0:M} | S_{0:M}) = p(Z_{0} | S_{0}) \prod_{i=m}^{M} p(Z_{i} | Z_{<i}, S_{\leq i}). \quad (6.9)
$$

Notably, the conditioning is applied to all states within the diffusion process i.e., for all $Z_{i}^{m}, t \in \{1, \ldots T\}$ at each diffusion step, we apply the cross attention as well as the memory attention module allowing us effectively capture the temporal context.

**Network Architecture.** To generate visual storylines, in our Story-LDM, we first introduce an autoregressive structure and modify the two-dimensional U-Net in [265] to so as to process the temporal information in the storyline. As shown in Fig. 6.2a, the frame encoder, $E$ equipped with the positional information of the frame in the sequence, is applied to get the low-dimensional representation $Z_{m}^{n}$ for all the frames in a datapoint from $\mathcal{D}$. A text-based transformer is applied to get a suitable representation for the sentence $S^{m}$. The U-Net is then applied to model the diffusion process over $T$ time steps. The layers within the U-Net are augmented with the cross-attention layer and our memory attention layer. After each downsampling or upsampling operation, we apply the attention mechanism to reinforce the conditioning on the already encoded (learned) storyline up to the previous time step. For any frame $m$, the cross-attention $C_{attn}$ is given by

$$
C_{attn} = \sum_{i} \hat{f}(Z_{m}), \hat{f}(S^{m}) \quad (6.10)
$$
where $\hat{f}(Z^m)$ and $f(S^m)$ are the representations of the frame encoding $Z^m$ within the neural network and sentence $S^m$ respectively such that they have same dimensions. Similarly, the memory attention $M_{attn}$ is computed as,

$$M_{attn} = \sum_{k=1}^{m-1} \sum_i \hat{f}(Z^k)_i f(S^k)_i f(S^m)_i$$  \hspace{1cm} (6.11)

The output of the attention module is then computed as the aggregation, $C_{attn} + M_{attn}$.

Starting from the noise sample $Z^m_t$, the output of the reverse diffusion process $Z^m_0$ is reconstructed using the frame decoder $D$ to get the final image. Having outlined the details of our Story-LDM framework, we show through extensive experiments on the task of story generation, the effectiveness and the benefits of our powerful conditioning based on memory-attention.

### 6.3 Datasets and Evaluation Metrics

In this work, we formulate story generation with co-references to actors and backgrounds across frames.

**Datasets.** Since reference resolution has not been studied in story generation, to validate our approach on this much harder task, we construct the following datasets: (i) We take an existing story-generation dataset – FlintstonesSV [97], and modify the sentences by replacing the named entities (characters) with references where possible; including pronouns such as he, she, or they (cf. Fig. 6.1). This dataset contains 20132-training, 2071-validation and 2309-test stories with 7 main characters and 323 backgrounds. (ii) MUGEN [104] is a video dataset collected from the open-sourced platform game CoinRun [49]. The dataset is divided into 104,796-train and 11,802 test stories with 96 to 602 frames. We extend the MUGEN dataset by introducing two additional characters Lisa and Jhon (we rename Mugen to Tony). We construct stories of four frames and corresponding text, ensuring consistent co-referencing in the story; each story has 3 such references. Moreover, we augment the existing two backgrounds (Planet and Snow) with four additional backgrounds: Sand, Dirt, Grass and Stone. (iii) We also modify the existing PororoSV [186] dataset which contains 10191/2334/2208 train/val/test set. Similarly, we reference characters by pro-nouns to generate a more natural story. We show in Fig. 6.1, example stories from the two modified datasets and enlist the complete statistics in Tab. 6.1.

**Evaluation Metrics.** To measure the consistency of the characters as well as the backgrounds in the generated stories, we consider the following evaluation metrics:


**Figure 6.5:** Qualitative Comparison on Story Generation. Comparison on the FlintstonesSV dataset visual story generation.

(i) **Character Classification:** Following [209], we consider fine-tuned Inception-v3 to measure the classification accuracy and F1-score. Frame accuracy evaluates the character match to the ground truth and F1-score measures the quality of generated characters in the predicted images. 

(ii) **Background Classification:** Similar to character classification, we use fine-tuned Inception-v3 to measure the correspondence of the background to the ground truth and consider the F1-score as a measure of quality. 

(iii) **Frechet Inception Distance (FID):** To assess the quality of images, we consider FID score [113] which is the distance between feature vectors from real and generated images.
6.4 Experiments

In this section, we evaluate our Story-LDM approach for consistent story generation with reference resolution.
Chapter 6. Visual Memory Conditioned Consistent Story Generation

Fred jumps over a stool in a room.

He is running across the room wearing a superhero costume and jumps over a chair.

Wilma walks from one room to another in the house.

Fred in a superhero outfit and Wilma are standing in the living room. Wilma talks to him while her arms folded in front.

Figure 6.7: **Story Diversity.** Diverse outputs for a single storyline obtained with our Story-LDM.

**Baselines.** We construct a strong baseline with the LDM\(^3\) [265] which contains a cross-attention layer to generate text-to-image based story, without using our proposed autoregressive memory modules as our baselines for MUGEN, Poro-roSV and FlintstonesSV datasets. The parameters of the diffusion model within the Story-LDM are initialized with the pre-trained LDM [265]. Similarly, for the textual embedding, we use BERT-tokenizer [58] and use the pre-trained text-transformer from LDM.

**Quantitative Results.** Table 6.2 shows quantitative results for consistent story generation on the FlintstoneSV dataset. We compare the performance of our approach (row 4) to the LDM [265] which we train/test with both the original (row 2) as well as the co-referenced (row 3) descriptions. Furthermore, we include the

\(^3\)https://github.com/CompVis/latent-diffusion
results of the state-of-art VLCStoryGAN [209] (row 1) with the original text of the dataset \(^4\) (i.e. without co-references). We note that VLCStoryGAN was shown to be better than Duco-StoryGAN [210], CP-CSV [289], and original StoryGAN [186] (see [209]).

Based on Table 6.2 we make three observations: (1) Our LDM baseline is better than VLCStoryGAN on the original reference-free text (cf. Tab. 6.2, rows 1 & 2). (2) Reference resolution makes the task considerably harder. With the reference text in our modified dataset, we observe a drop in performance in terms of character and background classification scores (cf. Tab. 6.2, rows 2 & 3). (3) Our model, with memory-attention module, significantly outperforms the baseline (cf. Tab. 6.2, rows 3 & 4) both in terms of generative image quality and character consistency; and outperforms SoTA of VLCStoryGAN by \(\sim 41\%\) percentage points on character accuracy (while performing a more difficult version of the task).

Further, our model, which is required to conduct reference resolution, comes close to the LDM trained with original, reference-free, text (cf. Tab. 6.2, rows 3 & 4), which can be viewed as a sort of an upper bound.

On the MUGEN dataset, our method outperforms the strong LDM baseline with gains of \(\sim 62\%\) on character accuracy and \(\sim 76\%\) on the background accuracy, thereby showing the advantages of the memory-attention mechanism for consistent story generation. We note that MUGEN dataset has more references across story scenes. Flintstones while contains more references per story overall, many of those references are within scenes as opposed to across scenes. Meaning that in terms of reference impact on consistency, MUGEN dataset is actually harder. Experimental results on the PororoSV dataset are provided in the Supplemental.

**Qualitative Results.** Figure 6.3 illustrates qualitative results on the MUGEN dataset. Rows 1, 2 & 3 show ground truth, LDM [265] and our Story-LDM approach, respectively. Here, we see that our method is able to maintain consistency in terms of both character and background. Similarly, in Figure 6.4 we can show the results on FlintstoneSV dataset which further validates the strong performance of our method when generating a high-quality, consistent story. Compared to the LDM, our approach is able to adapt to the diverse backgrounds in the story descriptions.

**Additional Results.** We compare the qualitative results of our method to both story generation [209] and story continuation [211] in Figs. 6.5 and 6.6 respectively. The comparative images are taken directly from respective papers. We note that story continuation Fig. 6.6 is solving a different (easier) problem and with text

\(^4\)Results for [209] were obtained using pre-trained model provided by original authors in private communication.
Chapter 6. Visual Memory Conditioned Consistent Story Generation

Figure 6.8: **Branching Storyline.** Generating different yet consistent stories by branching the storyline. Frames in the later columns are generated based on earlier ones and corresponding text.

that contains no references. This makes the comparison to our method, which receives fewer inputs, not very meaningful. Nether-the-less, our approach, which can resolve references and solve a harder story generation task, obtains highly competitive results.

Furthermore, to show that our autoregressive visual memory module can generate diverse stories conditioned on the current and previous condition, we create different storylines starting for a single sentence. In Fig. 6.8, we can see for reference ‘they’, the model can generate both the characters according to the storyline already parsed. Moreover, in Fig. 6.7 we show that our approach can not only generate consistent visual stories, but also diverse frames for the same text (cf. Fig. 6.7). Additional results are provided in the Supplemental.
Chapter 6. Visual Memory Conditioned Consistent Story Generation

Figure 6.9: Combination of different characters and backgrounds from the MUGEN dataset [104].

<table>
<thead>
<tr>
<th>Method</th>
<th>Reference-text</th>
<th>Char-acc (↑)</th>
<th>Char-F1 (↑)</th>
<th>BG-acc (↑)</th>
<th>BG-F1 (↑)</th>
<th>FID (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Story-LDM (Ours)</td>
<td>✓</td>
<td>69.19</td>
<td>86.59</td>
<td>35.21</td>
<td>28.80</td>
<td>69.49</td>
</tr>
<tr>
<td>Story-LDM (Ours)</td>
<td>×</td>
<td>83.29</td>
<td>94.61</td>
<td>35.54</td>
<td>27.32</td>
<td>64.89</td>
</tr>
</tbody>
</table>

Table 6.3: Experimental results using our proposed Story-LDM with and without the reference text on the FlintstonesSV dataset.

Table 6.4: Quantitative results. Experimental results on the PororoSV datasets.

### 6.5 Additional Quantitative Results

In Tab. 6.3, we show the performance of our Story-LDM approach with and without the reference text. We observe that without reference text *i.e.*, with explicit mentions of the characters (by their name), our approach outperforms the prior state-of-the-art VLCStoryGAN [210] (*cf.* Tab. 2, row 1) and the strong LDM [265] baseline on character accuracy by nearly $\sim 3.4\%$. 
This demonstrates that the introduction of our Memory attention module to the pipeline of the diffusion model with U-Net architecture increases the performance even in the traditional dataset setting. This can be attributed to the fact that even when the textual resolution of character names or settings is unnecessary, our memory attention module can still enhance the consistency of appearance in the visual domain. When using the descriptions with references, the character accuracy of the model drops by $\sim 12\%$ showing the difficulty of the extended task of generating stories from the co-referenced text. Even the strong LDM baseline which outperforms the prior state-of-the-art for story generation, offers limited performance on the complex task of story generation with co-reference resolution illustrating the hardness and complexity of the new task (cf. Tab. 2). Our Story-LDM approach outperforms this strong baseline illustrating the benefits of our autoregressive Story-LDM with memory module.

Table 6.5 shows quantitative results for consistent story generation with the reference text on the PororoSV [186] dataset. We compare our proposed Story-LDM with DUCO-STORYGAN [210] and VLCStoryGAN [209]. Our method outperforms previous baseline models (including LDM baseline) in character evaluation metrics. To be noted, PororoSV dataset has no background information, therefore we only perform a character-level evaluation on this dataset.

Moreover, we conduct human evaluation. We randomly select 10 examples to compare Story-LDM with LDM. Following [209], we select three evaluation criteria: visual quality, consistency, and relevance for each sample. We conducted a forced-choice experiment with 13 subjects. The preference rate for our Story-
LDM is 74.7%, 65.4% & 69.2% in terms of the listed criteria.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>w/ ref. text</th>
<th>Char-acc (↑)</th>
<th>Char-F1 (↑)</th>
<th>BG-acc (↑)</th>
<th>BG-F1 (↑)</th>
<th>FID (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flintstones</td>
<td>VLCStoryGAN [209] ×</td>
<td>27.73</td>
<td>42.01</td>
<td>4.83</td>
<td>16.49</td>
<td>120.85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LDM [265] ×</td>
<td>79.86</td>
<td>92.33</td>
<td>48.02</td>
<td>37.86</td>
<td>61.40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Story-LDM (Ours) ×</td>
<td>83.29</td>
<td>94.61</td>
<td>35.54</td>
<td>27.32</td>
<td>64.89</td>
<td></td>
</tr>
<tr>
<td>MUGEN</td>
<td>LDM [265] ×</td>
<td>95.25</td>
<td>97.04</td>
<td>21.10</td>
<td>23.98</td>
<td>123.69</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Story-LDM (Ours) ×</td>
<td>97.60</td>
<td>98.44</td>
<td>74.72</td>
<td>80.51</td>
<td>79.41</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.5: Experimental results using non-referential text on the Flintstones and Mugen dataset.

### 6.6 Additional Qualitative Results

In Fig. 6.9 we provide example image frames from our extended MUGEN dataset with three characters Lisa, Tony and Jhon, and six different backgrounds. Our extended MUGEN dataset is thus more complex than the MUGEN dataset in [104] where only one character is considered for two backgrounds.

Fig. 6.10 shows examples of positive/negative samples. To evaluate fore- and background consistency we propose two evaluation metrics: character and background classification. The generated story is considered accurate (positive sample) if it appropriately resolves character and background references in the generation, otherwise it is considered a negative sample. In the figure, the 1st row (i.e. positive sample) has 100% character and background accuracy; the 2nd row (i.e. negative) has 75% and 50% character and background accuracy.

Figs. 6.11, 6.12 and 6.13 show random samples obtained for story generation with our Story-LDM approach on the MUGEN, FlintstonesSV, and PororoSV datasets respectively. Clearly, our approach yields high-quality frames with character and background consistency.

### 6.7 Conclusion

In this work, we formulate consistent story generation in a more realistic way by co-referencing actors/backgrounds in the story descriptions. We develop an autoregressive Story-LDM approach with memory attention capable of maintaining consistency across the frames based on the previously generated frames and their corresponding descriptions. We introduced modified datasets to evaluate the performance for reference resolution. We expect our proposed formulation and models
Figure 6.11: Qualitative results on the MUGEN dataset [104].
to be conductive to the real-world use cases and further the research.
Pororo visits Loopy to sing for Loopy. Loopy is mad at Pororo. He is wearing a pig mask. He says sorry with his pig mask. Petty also said good bye to Poby and Harry. Petty was holding a piece of pie. Petty tried her pie. While Poby and Harry going back to their house, they heard a voice calling them behind.

Pororo and friends came to Poby and Harry. They were about to play hide and seek.

Poby understands Crong checks whether Crong wants one more story. Eddy raises his hand to express his opinion. He suggests that they read one new book. Eddy’s book is about dinosaurs. He holds the book by his hands and smile with it. Harry says yeah and flies to get closer to TongTong. Harry thinks that Tongtong is satisfied so Harry tries to sing another song. He asks Harry to sing a song next time.

Poby and Poby’s friends find Harry. Poby runs to Harry. Eddy says something to Rody. Rody is approaching Eddy. Rody points to the dish Eddy has. Eddy also points it. Eddy once something to Rody. Rody is approaching Eddy. Rody is thinking about how to throw the ball. Eddy is getting ready to hit the ball. Rody is holding the ball. Pororo is preparing to catch the ball. Eddy is getting ready to throw the ball. Eddy is practicing the ball.

Rody is about to throw the ball. Pororo is getting ready to catch the ball. Eddy is practicing the ball very confidently. Pororo is preparing to catch the ball.

Figure 6.13: Qualitative results on the PororoSV dataset [186].
Chapter 7

Visual Concept Driven Image Generation with Text-to-Image Diffusion Model

7.1 Introduction

Photorealistic image synthesis, particularly with text-to-image diffusion models (e.g., Stable Diffusion [266]), has provided a great boost to content creation. Images of diverse and imaginative scenarios such as “A dog in space” can now be easily generated. While such approaches can faithfully generate many instances of the generic subject, e.g., a dog, in a variety of scenarios, e.g., space, they are limited when the goal is to render a specific subject, e.g., my Beagle dog Bailey with unique fur markings, in a specific context or background, e.g., next to my Tudor-style house. Further, the generative capabilities fall apart when modeling multi-object scenarios and interactions, e.g., me playing with Bailey next to my house. Since these cases are frequent and typical in visual content creation, such as stories or videos, it is useful to consider how to make existing text-to-image diffusion models more controllable and capable of generating multiple interacting user-specified concepts. To introduce additional controllability, for example, concerning subject appearance or style and their occurrence in different contexts, many recent approaches [78, 98, 270, 349] have focused on controlling generation with sample images depicting concepts of interest. For example, to render a specific subject, these approaches present multiple (usually relatively clean) images of the target subject to extract the single subject-specific token across these images, e.g., using inversion techniques, and then introduce this special token in prompts in conjunction with the regular text to generate desired content. However, while one can generate single concept/subject scenarios in this way, composing multiple stylistically diverse concepts remains a challenge.

Consider images in Fig. 7.1; approaches discussed above are capable of extracting the specific subject (from the left concept images) and “pasting” them in different contexts. However, they have shown to be significantly limited [16] when...
Chapter 7. Visual Concept Driven Image Generation with Text-to-Image Diffusion Model

Figure 7.1: **Concept-driven image generation:** Given images depicting multiple concepts (subjects and context/background), the top output is the illustration of the male ([v1]) being generated in the context/background of the female ([v3]) image by different methods. The bottom output illustrates the two concepts together in a single image. Dreambooth [270] (left) encodes [v1], [v2], [v3] from multiple input concept images. It fails to generate multi-concept interactions. Break-a-scene [16] (middle) disentangles [v1], [v2], [v3] from single image. This approach requires human-annotated masks. Our approach (right) disentangles [v1], [v2], [v3] from single image. The latent masks are obtained from EM-like optimization. Using these optimized masks, our method can automatically produce images with those concepts in new contexts either by themselves or, jointly, interacting with one another.

asked to generate images where, for example, two characters in question interact or when multiple concepts come from the same image (e.g., the top-left image can be the source of both the subject and the background). Learning of tokens and subsequent generation, in this scenario, requires explicit disentanglement of the characters and the background. Further, stylistic variations (top – game rendering; bottom – real photograph) present added challenges.

To address these limitations, recent work [16] proposes a two-stage training pipeline for extracting multiple concepts from a single image by leveraging a user-specified mask for each of the concepts during the optimization of tokens associated with these concepts. This mechanism enables the decomposition of an image into constituent subjects and the background. Analogous to prior methods that use multiple images to learn a single concept, this approach also learns how to encode features of each of the concepts into corresponding tokens. A key feature of the pipeline lies in the fine-tuning of model weights while optimizing the prompts...
for different concepts simultaneously. This ensures that multiple concepts can be compositionally combined. While this works well in practice, the process requires users to specify masks for concepts they desire, which requires a custom sketching interface and additional user effort (as shown in Fig. 7.1, middle).

In this work, we attempt to do away with the need to specify masks for multi-concept extraction and composition, while still leveraging (latent) masking to learn disentangled concepts that can then be composed using the text-to-image diffusion model [266]. Specifically, we utilize loose textual input prompts and the cross-attention maps for the target concept tokens to generate masks during prompt optimization. However, generating concept masks from cross-attention (e.g., as is done in [337]) of pre-trained diffusion model tends to be noisy; often under- or over-segmenting the target concept(s). Therefore, we propose a joint EM-style optimization over the concept tokens and their (latent) masks, where given mask initializations (obtain in the style of [337] from cross-attention) we iterate between learning tokens for masked concepts and re-estimation of the mask itself. We illustrate that this leads to both learning of better concept tokens and, as a by-product, a better latent mask for the concepts in question.

As is illustrated in Fig. 7.1 the proposed approach can automatically learn tokens corresponding to the two subjects ([v1], [v2]) and their backgrounds. Given these tokens, it can seamlessly integrate subject in [v3] within the environment of [v2] and can also generate plausible interactions between the two subjects. Note that it can do so despite stylistic differences between the concept sources.

Our contributions are:

• We address a unique and challenging problem of personalized multi-subject generation with complex interactions of (potentially unrelated) subjects.

• We propose an Expectation Maximization (EM)-like optimization procedure to disentangle concepts from a single, or multiple, images by generating masks for specific target concepts. Our approach performs token optimization while simultaneously optimizing corresponding (latent) masks such that the token embedding best represents the disentangled concept.

• We show through novel combinations of subjects and interactive environments that our approach works well in practice and can generate realistic scenarios with custom subjects interacting in complex scenes and environments.
7.2 Related Work

Image and text-to-image generation. Realistic and high-quality image synthesis is now possible with the outstanding advances in deep generative models. Generative Adversarial Networks (GANs) [1–3, 34, 78, 89, 144, 145, 177] facilitate the creation of high-fidelity images spanning different domains. In addition to GANs, Variational Autoencoders (VAEs) [154, 314] have presented a likelihood-based method for generating images. Important class of generative models include autoregressive models [43, 68, 238, 315] and diffusion models [38, 59, 114, 115, 224, 266, 271]. The former views image pixels as a sequence with pixel-by-pixel correlation, while the latter generates images by gradually removing noise. Methods such as DALL-E [260] employed an autoregressive transformer [317] trained on both text and image tokens, showcasing remarkable zero-shot capabilities.

Diffusion-based models [114] with their spectacular realistic synthesis capa-
Chapter 7. Visual Concept Driven Image Generation with Text-to-Image Diffusion Model

Personalizing image generation. Personalization involves identifying a concept provided by the user that is not commonly found in the training data for discriminative [50] or generative [226] purposes. Current approaches for customized text-to-image generation with diffusion models have focussed on two types of strategies. In the first, methods [40, 72, 99, 170, 205, 270] adopt test-time fine-tuning where model weights are updated using a collection of images representative of the target concept/subject for personalization. The second strategy is to focus solely on refining the token embedding representing the subject while keeping the model weights fixed to enhance the model’s comprehension of visual concepts [78, 98, 349]. In this direction, DreamBooth [270] takes a holistic approach by fine-tuning the entirety of the UNet network. In contrast, Custom Diffusion [170] concentrates its fine-tuning efforts exclusively on the K and V layers within the UNet network’s cross-attention mechanism. This fine-tuning process is then optimized using LoRA [125]. Conversely, SVDiff [99] pioneers the use of cut mix to construct training data and introduces regularization penalties to mitigate overlapping attention maps associated with multiple subjects during the training process. Cones [197] introduces the notion of concept neurons and exclusively modifies these neurons related to a specific subject within the K and V layers of cross-attention. To generate multiple personalized subjects, Cones directly combines the concept neurons from several trained personalized models. In contrast, Mix-of-Show [94] adopts a different approach by training individual LoRA models for each subject and subsequently merging their outputs through fusion techniques. In this work, we explore the integration of subjects extracted from multiple images, examining their mutual influence to generate images driven by these subjects. We utilize cross-attention maps to disentangle acquired subjects by optimizing masks.

Inversion in generative models. Inversion tasks in generative models involve identifying a latent code, typically within the latent space of a generative model,
that can faithfully recreate a specific image \([340, 379]\). The process of inversion is achieved either through optimization-based methods that directly refine a latent vector \([1, 92, 383]\) or by employing an encoder to discern and generate the latent representation corresponding to a given image \([264, 309, 378]\). In this work, we adhere to the optimization method, considering its superior adaptability to novel or unfamiliar concepts. The work incorporates a dual-phase strategy. Initially, we optimize solely the textual embeddings related to the target concepts using masks generated by cross-attention maps. Subsequently, we undertake joint training, refining both the embeddings through mask optimization and the model weights simultaneously.

### 7.3 Methodology

To generate images with user-specified concepts (subjects or background) and enable interactions between multiple such concepts and in various contexts, it is essential to extract the concept information, disentangle its specific visual representation, and encode it in the diffusion model. We define a *concept* as a visually illustrated instance of an object/subject or a well-exemplified pattern. Note that such concepts often entangle content and style.

Diffusion models encode the textual guidance for accurate image generation through cross-attention between the textual tokens and the corresponding region in the image. Therefore, to associate *new* tokens with particular concepts, a masked representation of the target image highlighting only the specific concept is inevitably beneficial. We consider an alternating expectation maximization (EM) style optimization to jointly learn: the tokens – encoding the concept-specific information; and the binary masks – corresponding to each concept of interest. This methodology provides for accurate reconstruction of the target concept in arbitrary contexts, even when involving complex interactions between the subjects. In the following, we first introduce the backbone and the constituent components of our optimization framework.

**Latent diffusion model.** Latent Diffusion Models (LDM) \([266]\) use perceptual image compression to project the original image to a lower-dimensional space. An encoder \(E(\cdot)\) maps the input image \(I \in \mathbb{R}^{H \times W \times 3}\) to a latent representation \(Z \in \mathbb{R}^{h \times w \times c}\), downsampling the image to a lower spatial dimension. The diffusion model is then applied to the latent \(Z\), where a time-conditioned U-Net \(\epsilon_\theta(Z_t, t)\) models the diffusion process. The objective of the latent diffusion model is,

\[
\mathcal{L}_{LDM} := E_{t, Z, \epsilon} \left[ \| \epsilon - \epsilon_\theta(Z_t, t) \|^2_2 \right].
\]
Figure 7.3: **Quantitative comparison - Mask IoU as a function of training steps.**

We compare the estimated (latent) mask and the manually annotated mask for the concepts illustrated in Figure 7.1. Performance, as a function of mask optimization, for different tokens over training steps is illustrated: (a) Mask optimization for Token1 (*i.e.* subject), (b) Mask optimization for Token2 (*i.e.* another subject), and (c) Mask optimization for background Token. The yellow line indicates a baseline where the mask is obtained once and not updated jointly with the tokens (hence fixed value). A clear trend of improvement is observed for all masks. Curves are smoothed to highlight trends.
Mask extraction. Following [337], we leverage the cross-attention maps to generate the mask of the target concept. Given a conditioning prompt $y \in \mathbb{R}^{N \times d}$ where $N$ is the number of tokens in the prompt and $d$ is the embedding dimension, a cross-attention layer with key ($K$), query ($Q$) and value ($V$) is given by,

$$A = \text{softmax} \left( \frac{QK^T}{\sqrt{d}} \right) \cdot V. \quad (7.2)$$

The attention map $A^1$ can be reshaped to $H \times W \times N$ such that each slice $A_p \in \mathbb{R}^{H \times W}$ represents the region attended by the $p^{th}$ token. These attention maps are averaged across all layers of the U-Net to get the response for the regions attended by the target tokens. Note that the latent diffusion objective and hence the attention map are a function of the timestep $t$. Therefore the quality of the mask would in general differ depending on one or a range of timesteps considered. We will discuss this in more detail later.

To obtain the binary mask $B$, we follow the simple process of using a fixed threshold and assigning a value of 1 for pixel values larger than the threshold and 0 otherwise. We further refine the binary masks using the dense conditional random field (DenseCRF) [161], where a DenseCRF, or fully connected CRF, endows pairwise potentials for each pair of pixels in an image where the local relationship, which is based on the color and distance of pixels. The goal of CRF optimization is to satisfy the soft unary constraints obtained by passing a stack of binary masks for each image through a softmax, while attempting to ensure nearby and similar pixels are assigned the same class label. It yields refined segmentation that respects natural segment boundaries in the image, which we call $M$. We adopt DenseCRF hyperparameters from CutLER [332].

Masked diffusion loss. The input prompt for token optimization is designed such that it contains tokens for target concepts. In a prompt $y$ with $k$ tokens of interest, diffusion loss is applied to the corresponding pixels which in turn are obtained from the binary mask of each token. Therefore, considering $M_k = \bigcup_{i=1}^k M_i$ to the union of the pixels of $k$ concepts, the masked diffusion loss [16] is given by,

$$L_{Mask} := \mathbb{E}_{t, Z, \epsilon} \left[ \| \epsilon \odot M_k - \epsilon_{\theta}(Z_t, t) \odot M_k \|_2^2 \right]. \quad (7.3)$$

Cross attention loss. The masked diffusion loss attends to all the desired concept tokens in a prompt. To enforce that a single token encodes information specific

---

1Here, $Q = W_Q \cdot f(Z)$, $K = W_K \cdot f(y)$ and $V = W_V \cdot f(y)$, and $W_Q \in \mathbb{R}^{d \times d_q}$, $W_K \in \mathbb{R}^{d \times d_k}$ and $W_V \in \mathbb{R}^{d \times d_v}$ are learnable parameters, $f(Z)$ an intermediate flattened feature representation of $Z$ within the diffusion model and $f(y)$ the feature representation of the condition $y$. 
to its corresponding instance, we also include the cross-attention loss [16] which encourages the token to attend to only the corresponding target concept,

\[ L_{\text{attn}} := \mathbb{E}_{i, t} \left[ \| C_{\text{attn}}(Z_t, y_i) - M_i \|_2^2 \right]. \quad (7.4) \]

Here, \( C_{\text{attn}}(\cdot) \) is the cross-attention map between the visual representation \( Z_t \) at timestep \( t \) and the token \( y_i \) in prompt \( y \) with \( k \) concepts, and \( M_i \) is the (latent) mask for token \( y_i \).

**Optimization.** To ensure that the tokens learned correspond to the target concept, we introduce binary masks to constrain the loss during prompt optimization and update the tokens under the mask. We adopt a weakly supervised approach and an alternating optimization procedure for learning the concepts. An EM optimization leads to better convergence under good initialization of the parameters.

To obtain a reasonable initial estimate of the token embeddings of specific concept-tokens and their corresponding masks, in the first stage, we perform the following updates: We first obtain initial concept token embeddings by either initializing them from CLIP or by performing inversion without masking to optimize the tokens for a fixed number of iterations. This provides us with good initial token embeddings. With these initial token embeddings at hand, we initialize the masks by following the steps outlined in mask extraction and average the masks over certain timesteps to reduce the errors introduced by stochasticity.

In the second stage, given good initializations of the tokens and their masks, we perform an alternating update of the tokens given the masks and of the masks themselves. Using Eq. (7.3) and Eq. (7.4), we first optimize the prompt given the masks for each concept at the iteration. In the following step, for the optimized token, we compute the Eq. (7.3) to update the mask (from cross-attention). Finally, we optimize the tokens and parameters of the model jointly keeping the mask fixed, similar to [16], this avoids drift.

The overall optimization is outlined in Fig. 7.2. In this example, tokens \([v1]\), \([v2]\), and \([v3]\) are assigned to the two subjects and a background of the second image respectively. With a good initialization of the token embeddings, we obtain the initial masks of the three concepts by averaging the attention maps over certain timesteps. Following this, we apply \( L_{\text{Mask}} \) and \( L_{\text{attn}} \) to update the tokens constrained by the given masks. With the newly updated tokens, we update the model weights to get an updated, better-quality mask. We repeat this process over a fixed number of steps (usually 400). At each iteration, we sample the timesteps to use for the update, similar to [16].

In this way, our approach can disentangle and learn concepts from a single image with masks generated during optimization. As illustrated in Fig. 7.3, the IoU of
Figure 7.4: **Qualitative results for concept-driven image generation.** Here, we show a comparison between our results with mask optimization, baseline method without mask optimization, and ground truth approach.

Our optimized masks with respect to the ground-truth masks consistently increases for different concepts compared to the masks obtained from baseline (without optimization). Our approach thus provides flexibility to extract complex concepts without the requirement of a user-defined ground-truth mask.

**Algorithm** We show an overview of our proposed visual concept-driven image-generation approach in Algorithm 1. The algorithm presents an optimization process akin to Expectation Maximization (EM) to untangle different elements within images. It achieves this by creating masks that isolate particular desired concepts within the images, while, at the same time, learning tokens to represent them. This process of optimization not only improves the learning of more effective concept tokens but also yields enhanced latent masks for specific visual concepts.
Chapter 7. Visual Concept Driven Image Generation with Text-to-Image Diffusion Model

Algorithm 1: EM Optimization of Prompt Tokens and Masks

1. Input: Diffusion model parameters: $\theta$, Input image: $Z = E(I)$, Initial prompt: $y$, Prompt embedding: $\hat{e}$, Target concept indices: $K$; Learning rate: $\lambda$, Timesteps for attention and masks: $T$, Optimization steps: $N$

   /* Sample $C$ timesteps */
   $T \leftarrow \text{random}(0, T, C)$

   /* Prompt Initialization */
   for $p \in K$
   for $t \in T$
   $g = \nabla_{\hat{e}_p} [\mathcal{L}_{LDI}(\epsilon_{\theta}(Z_t, t, \hat{e}))]$
   $\hat{e}_p = \hat{e}_p - \lambda g$
   end
   end
   /* Initial Attention Map */
   $A_p \leftarrow \text{From ???}$
   /* Mask Initialization */
   for $p \in K$
   $B_p \leftarrow \frac{1}{|T|} \sum_t \text{binarize}((A_p)_t)$
   $M_p \leftarrow \text{DenseCRF}(B_p)$
   end
   /* EM optimization of tokens and masks */
   for $n \leftarrow 1$ to $N$
   /* Sample $C$ timesteps */
   $T \leftarrow \text{random}(0, T, C)$
   /* Prompt optimization */
   for $p \in K$
   for $t \in T$
   $g = \nabla_{\hat{e}_p} [\mathcal{L}_{\text{mask}}(\epsilon_{\theta}(Z_t, t, \hat{e}), M_p)]$
   $\hat{e}_p = \hat{e}_p - \lambda g$
   end
   end
   /* Mask optimization */
   for $p \in K$
   for $t \in T$
   $g = \nabla_{\theta} [\mathcal{L}_{\text{attn}}(C_{\text{attn}}(Z_t, t, \hat{e}_p), M_p)]$
   $\theta = \theta - \lambda g$
   end
   end
   $A_p \leftarrow \text{Update using Eq. (7.2)}$
   $B_p \leftarrow \frac{1}{|T|} \sum_t \text{binarize}((A_p)_t)$
   $M_p \leftarrow \text{DenseCRF}(B_p)$
   end
   return $\hat{e}$

7.4 Experimental Results

In this section, we will show the experimental results through both quantitative measurements and qualitative assessments. Additionally, we conduct comparisons with baselines for a comprehensive evaluation.

Baseline Methods. We consider two baselines for comparing our proposed subject-
driven image generation method, both of which are subsets of the Break-A-Scene [16]. These subsets involve: (1) utilizing user-defined masks (i.e. ground truth masks) and (2) operating without user-defined masks, which is a weakly supervised baseline (baseline method) to compare with our method. In this specific baseline approach, we create initial masks by aggregating attention maps derived from a pre-trained stable diffusion [266] model using 50 timesteps, rather than relying on ground truth masks. However, there is no optimization of these initial masks in the baseline method. We note that (1), which uses human-annotated perfect masks to delineate concepts, is a natural upper bound for our method. That said, it is possible in certain cases for us to outperform it as full-mask may not be optimal for learning the concept in question; the learned (latent) mask allows added flexibility in this respect.

Quantitative Comparison. We consider three evaluation metrics to evaluate the faithfulness of the concept-driven generated images approach to the input concepts,
Table 7.2: Quantitative evaluation based on user study.

<table>
<thead>
<tr>
<th>Data</th>
<th>Baseline</th>
<th>Neutral</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>10.92%</td>
<td>17.29%</td>
<td>71.80%</td>
</tr>
<tr>
<td>COCO</td>
<td>21.87%</td>
<td>18.69%</td>
<td>59.44%</td>
</tr>
</tbody>
</table>

considering two measurement criteria as follows:

- Similarity of the generated images: We use CLIP [247] and Cosine Similarity [171] scores to evaluate the resemblance of the generated images to the target images in terms of semantic similarity. Given a generated image, we use a pre-trained object detector (e.g., YOLOv8 [131]) to detect the object of interest (corresponding to the input concept) in the generated image and compare their similarity to the target object from MSCOCO images. Table 7.1 displays the results for the baseline method and our method using CLIP and cosine similarity.

- Diversity between the generated images: We also consider how diverse the generated images for visual concept-driven generation are for a given prompt by comparing them using LPIPS [364] across different samples in the Table 7.1.

In Table 7.1, we present results on the COCO dataset [191]. We randomly selected 120 concepts from 11 COCO object classes, and generated them in various settings using 20 prompts each, creating 8 images for each prompt. We consider our approach with and without data augmentation. Data augmentation makes the generated images more diverse, but it can lead to lower similarity scores when compared to the target images. To account for this, similar to the baseline, we also evaluate our method without data augmentation and calculate semantic similarity using CLIP and cosine similarity scores. Our method without augmentation still performs better in diversity, due to EM-like optimization, allowing it to generate a wider variety of images compared to the baseline. This shows the importance of mask optimization in the process of generating masks.

Moreover, we also found human evaluation to be the best possible way to validate the approach. We randomly select 10 prompts and generate 6 images for each of them. We carried out a forced-choice experiment involving 11 participants to compare the generated images between our method and the baseline approach.

These images were assessed for faithfulness to the target instance and the prompt condition (representing instances, contexts, and interactions) used to gen-
Figure 7.6: **Qualitative results demonstrating the interaction between different concepts.** Here we present the interaction between newly learned concepts and the pre-existing concepts already familiar to the stable diffusion model.

generate the images. As shown in Table 7.2, our method was preferred 71.8% of the time and was considered on par 17.29%. In addition, we conducted another, much larger scale one, using COCO in the same setting as in Tab. 7.1. In a user study with 17 participants, our approach was chosen 59.44% of the time, compared to the baseline. This clearly illustrates the preference for results obtained using the joint token-mask optimization we propose.

**Qualitative Comparisons.** Figure 7.4 shows the results of concept-driven image generation using our method, ground-truth (GT) [16], and the baseline method. The bottom rows exhibit the ground-truth masks, the respective learned concepts from two different use cases, and the images generated for interactions with multiple concepts. The middle rows and the top rows showcase the weakly supervised generated masks using the baseline and our method, respectively. In comparison to the baseline, our approach yields images that faithfully represent each concept in multi-concept image generation.

Figure 7.5 shows qualitative comparisons between the user-defined ground truth masks and the masks generated by both the baseline method and our approach. We can see that our optimized masks outperform the baseline masks without any optimization.
Figure 7.7: **Diverse images for the same prompt.** Given input concepts, our method demonstrates the ability to generate a variety of diverse outputs for a single prompt.

Figure 7.8: **Interaction between cartoon or real concepts (including background).** Here, our method is not only able to interact with newly learned concepts but also engage with pre-existing concepts (such as Obama and Trump) in a cartoon-like appearance.
Figure 7.9: **Comparison of our method with standard baselines.** Here we consider TI-m [78], DB-m [270], CD-m [170], ELITE [334], Break-a-scene [16] as standard baseline methods and all images are taken directly from [16].

Figure 7.10: **Qualitative results show that our model can handle four concepts simultaneously.**

Figure 7.11: Qualitative ablation by removing different components.
Chapter 7. Visual Concept Driven Image Generation with Text-to-Image Diffusion Model

Figure 7.12: **Interaction between real-world concepts (including background).** Our model demonstrates the capacity to generate accurate interactions among real-world concepts.

**Additional Results.** In Figure 7.6, we present the comparative results to show the interaction between newly introduced concepts and the pre-existing ones within the stable diffusion model. Our proposed method is not only able to interact between multiple concepts but also shows its capability to produce diverse outputs for a single prompt, as illustrated in Figure 7.7.

Furthermore, we demonstrate (in Figure 7.8) that our approach can interact within a cartoon world context as well. In this scenario, the pre-existing characters (e.g. Obama and Trump) are also depicted in a cartoon-like appearance. Besides, apart from lexical and cartoon concepts, we conduct experiments utilizing real-world concepts and show that our approach can generate multi-concept interactions between pre-existing and new concepts. We compare the qualitative results of our method in Figure 7.9. The comparative images are taken directly from
Figure 7.13: **Generated mask w/ and w/o DenseCRF.**

Break-a-Scene [16]. We can see that our method can generate a better-optimized mask without any spatial supervision. Additionally, the generated images are more aligned with specific subjects and adhere more closely to the provided text prompts. 

As an additional experiment, we conduct experiments to see if our model can learn more than three concepts (e.g., four concepts), and the qualitative results are shown in Figure 7.10. We can see that our method excels in learning individual concepts but encounters challenges in effectively capturing interactions between these concepts. A possible extension of our proposed approach could address this limitation, and improve the approach to capture interactions between different concepts.

In Figure 7.12, we also present qualitative results that demonstrate our model’s ability to engage with real-world concepts. Moreover, we show the importance of adding DenseCRF refinement while generating masks in figure 7.13. Looking at the figure, it is evident that the DenseCRF refinement significantly enhances the fine details and boundaries in the resulting masks. Refining the initial predictions and taking into account pixel-level interactions and dependencies, leads to a more accurate and visually appealing segmentation mask. In addition, we also present qualitative results where we learn about both single and multi-concepts from the
COCO [191] images in figures 7.14 and 7.15 respectively. We also compare the results generated by our method with the baseline method. From the figures, it is clearly visible that our method is able to maintain fine details about the visual concept compared to the baseline method.

Figure 7.14: **Qualitative results for single concept image generation.** Here, we show a comparison between our method and baseline approach using COCO images.

**Ablation Studies.** To assess the contribution of each component, we conduct qualitative ablation by removing DenseCRF refinement, cross-attention loss, and masked diffusion loss in Figure 7.11. We can see that when we remove one of the components, the model tends to entangle concepts. Similar behavior has been observed in [16] for masked diffusion and cross-attention loss. Moreover, adding DenseCRF refinement improves the generated masks by 5.9% in terms of IOU.

**Runtime cost:** The additional cost during training is just 0.5% more than the baseline method and 0.63% more than the break-a-scene. The inference cost remains the same across all models.
Figure 7.15: **Qualitative results for multi-concept image generation.** Here, we show a comparison between our method and baseline approach with interactions where input images are taken from COCO dataset.

Figure 7.16: **Utilization of multiple images for a single concept.** The model can grasp individual concepts well but faces difficulty in connecting or interacting between these concepts.
Chapter 7. Visual Concept Driven Image Generation with Text-to-Image Diffusion Model

Figure 7.17: **Stable Diffusion Age Bias.** The model demonstrates a bias towards adult individuals, resulting in a noticeable alteration in the boy’s appearance.

Figure 7.18: **Stable Diffusion Caucasian Bias.** The model exhibits a stronger bias towards data associated with the Caucasian demographic or region.

**Why EM-like optimization:** In our proposed approach, we follow a cyclical process reminiscent of the Expectation-Maximization (EM) algorithm. The analogy to the EM algorithm implies that the alternating steps contribute to refining the overall optimization process. In EM, the E-step involves estimating the expected values of the hidden variables given the observed data, and the M-step involves maximizing the likelihood of the parameters, incorporating the information from the E-step. Similarly, in our method, the alternating steps likely contribute to improving the model’s understanding of both the custom tokens and latent masks, with each iteration refining the knowledge gained from the other.

### 7.5 Limitations and Future Directions

Our proposed approach and Break-a-Scene [16] both employ single images to learn specific concepts. However, considering the utilization of multiple images for a single concept, could this potentially enhance the learning of the concept? The qualitative results for the experiment are shown in Figure 7.16. We observed that multiple images for a single concept do not enhance the interaction between multiple concepts. Despite the model’s improved understanding of each concept, there is no substantial improvement in the overall interaction among concepts. Therefore,
it could be a potential future direction to improve the quality of generated images.

In this work, we adopted a pre-trained stable diffusion [98] model as the foundation and fine-tuned it to grasp concepts for improved image generation. However, our observations revealed the presence of stable biases, including age and Caucasian bias, as depicted in Figures 7.17 and 7.18 respectively. Working to remove these stable diffusion biases could be another potential research direction to extend the proposed approach.

7.6 Conclusion

We present an approach for generating multiple subjects with interconnections to facilitate personalized generation, even incorporating entities beyond out-of-distribution. Our method involves an alternating optimization method aimed at separating different concepts within a single image by producing masks tailored to distinct target concepts. We anticipate its role as a foundational element in the evolution of generative AI, further expanding the horizons of creative expression.
Chapter 8

Conclusions

8.1 Research summary and impact

Our research mainly focuses on addressing the question: "What" and "How" multimodal learning can achieve human-level perception. In this thesis, we explored the basic ideas of multimodal learning and framed our main research questions (RQ) around these concepts as follows:

- **RQ$_1$:** Which modalities play a crucial role in addressing vision-related issues, and can incorporating audio data provide additional information?

- **RQ$_2$:** Is it possible to extract the desired audio information from a mixture that includes noise?

- **RQ$_3$:** How can we develop effective learning approaches when dealing with limited multimodal data, specifically under conditions of weak supervision?

- **RQ$_4$:** Can we extract generic features by leveraging multiple modalities?

- **RQ$_5$:** How do we address data ambiguity when dealing with multiple modalities?

Each research question addressed in this thesis contributes significantly to the advancement of multimodal learning. Understanding the crucial modalities in addressing vision-related issues is fundamental for fields such as computer vision and healthcare, which is what the first research question (RQ$_1$) is all about. For instance, the integration of audio alongside visual features substantially enhances the performance of dense event captioning. Therefore, it influences the choice of modalities integrated into multimodal models. Incorporating audio data can potentially enhance information processing, leading to improved solutions for visually impaired individuals, surveillance, or medical diagnostics. However, audio may be blended with noise or a combination of other audio sources. Hence, **RQ$_2$ explores the feasibility of extracting relevant audio information from multiple audio mixtures.** This has significant implications for applications involving sound data, enhancing the reliability and effectiveness of audio-based features in multimodal
systems, such as speech recognition, audio surveillance, and communication in noisy environments. Moreover, RQ3 focuses on developing effective learning approaches when dealing with limited multimodal data under conditions of weak supervision. Strategies derived from this research question can include transfer learning, domain adaptation, and semi-supervised learning, addressing challenges associated with constrained data availability. This has applications in various fields such as healthcare diagnostics, where obtaining labeled data is often limited, and weakly supervised learning can be a valuable solution. The fourth research question (RQ4) investigates the extraction of generic features from multiple modalities. The goal is to identify recurring patterns or features that are present across multiple instances or types of data within a dataset. These features provide a versatile representation that is not specific to individual instances. Instead, they capture the underlying information shared among diverse data points. This exploration contributes to the development of versatile multimodal models capable of generalizing well across different tasks and domains. For example, consider a multimodal learning scenario where information comes from text and images for sentiment analysis in social media posts. By extracting and integrating these generic features, a multimodal model can effectively capture the nuanced expression of sentiment in social media posts, providing a more accurate and holistic understanding. The model can then generalize well to new posts, adapt to different social media platforms, and offer interpretable insights into the factors influencing sentiment predictions. Finally, RQ5 addresses the challenge of data ambiguity in multiple modalities, especially text-to-image generation. Methods derived from this research question aim to enhance the robustness of multimodal learning models, making them more reliable in real-world scenarios where data may be inconsistent or incomplete. Specifically, the developed techniques align the semantic information in textual descriptions with visual features, ensuring that the generated images accurately represent the intended content. Improved semantic alignment can lead to more coherent and contextually relevant text-to-image generation. This has applications in fields like advertising, design, and content creation, where accurate visual representation is essential.

In summary, addressing these research questions can contribute to advancements in various domains, including healthcare, surveillance, communication systems, robotics, and natural language processing. The impact lies in improving the effectiveness, robustness, and versatility of multimodal data processing and machine learning systems.
8.2 Extensions and perspectives of my research

Research in this thesis primarily centered around efficiently learning from and utilizing multimodal data, including images, audio, human poses, and text. It particularly aimed to address the complexities arising from multiple modalities and tackled issues related to ambiguity. Future research would likely, and there is already much evidence for this, including the work discussed in Chapters 6 and 7, focusing on multimodal generative AI.

As multimodal strategies increasingly integrate into artificial intelligence systems, extensions of this thesis may address the broader issues that arise in tandem with the escalating intricacy of data and the contextual complexity of deployment in various environments. For instance, consider the development of an autonomous driving system that incorporates various sensors such as cameras, LiDAR, and radar to perceive its surroundings. The research in the thesis aims to devise methods that effectively fuse information from these sensors, ensuring a coherent representation of the environment. This involves developing advanced models that can align the modalities, handle missing data, and adapt to varying environmental conditions. Furthermore, future research likely to focus on shared architectures capable of handling diverse modalities, as opposed to utilizing individual networks for each modality. This approach brings benefits in terms of computational efficiency and model consistency. Consider a health monitoring system that collects data from wearable devices, including heart rate readings and sleep patterns, along with user-generated text logs about their daily activities. Instead of designing distinct models for processing each type of data, future models would likely create a unified architecture that can comprehensively analyze all modalities. This not only simplifies system maintenance but also provides users with a cohesive and holistic view of their health and lifestyle.

There are other areas that may also be valuable to explore in the future. For instance, techniques that can model uncertainty in vision problems and across modalities involved, as well as methods for detecting instances that fall outside the data distribution using deep generative models.

Moreover, exploring biases in machine learning systems constitutes a foundational research area. Multimodal learning, which involves using different types of information, can have biases from each type, and there might be new biases when these types are combined (for example, relying too much on one type of information). Although research in this thesis did not explore such biases, this is an important direction that needs to be explored in the future.

Outcome and Significance. The successful culmination of this research holds significant value across various domains where tasks necessitate the integration of
multiple modes of input data or deal with incomplete modalities. In our daily lives, numerous prospects abound for leveraging multimodal deep learning. The intended outcome of this research is the creation of a prototype for multimodal deep learning that can address tasks associated with human-centric videos, mirroring our abilities to perceive and comprehend through all our senses. This prototype would not only advance the field of computer vision but also find practical implementation in an array of real-world scenarios, contributing to enhanced human-computer interaction and understanding.

8.3 Ethics Assessment

In this thesis, we primarily explore multimodal learning. Multimodal learning plays a pivotal role in many critical applications. For example, in self-driving cars, it helps to see and understand the road by combining data from cameras, sensors, and more. It makes driving safer and more comfortable. Similarly, in medicine, it helps doctors by putting together information from different scans like MRI or CT scans, making it easier to see and understand what’s going on in our bodies.

However, there are risks to be considered. In both autonomous driving and medical imaging, the risk of misclassifications is a critical concern associated with the implementation of multimodal learning. In the context of self-driving cars, misclassifications can occur when the system misinterprets environmental cues, leading to errors in recognizing road conditions, obstacles, or other vehicles. Adverse weather conditions further amplify this risk, as sensors may struggle to accurately interpret the environment during heavy rain or snow, potentially compromising the vehicle’s safety. Complex traffic scenarios pose additional challenges, where the system may misclassify diverse objects, resulting in inaccurate decision-making.

Similarly, in medical imaging applications, misclassifications can have severe consequences, potentially leading to incorrect diagnoses. The multimodal learning system may misinterpret imaging data, causing false positives or negatives and impacting the precision of disease detection and characterization. Therefore, addressing these risks requires continuous validation, robust testing, and the development of adaptive algorithms to ensure accurate classifications and mitigate the potential consequences of misclassifications in autonomous driving and medical imaging. Similar considerations must be made in wider applications of multimodal learning.

A part of this thesis also focuses on generative models, which create new content like images or text, and come with certain risks. Sometimes, the generated content may not look realistic or could have mistakes, making it less useful. These models can also learn and repeat biases from the data they were trained on, leading to unfair or discriminatory results. We have to consider techniques for detect-
ing and mitigating biases in both the training data and the generated content. This might involve preprocessing the data to remove biases or adjusting the model architecture to reduce bias amplification. Another risk is that generative models might accidentally, or intentionally, produce offensive or harmful content. Ensuring that generative models are used ethically and responsibly is crucial to prevent misuse, comply with regulations, and minimize potential harm. We can incorporate human oversight and review mechanisms to evaluate the quality and fairness of the generated content. Human reviewers can identify and flag offensive or harmful content produced by the model. We can also apply post-processing filters or algorithms to the generated content to identify and remove offensive or harmful material. This can include language filters or image recognition algorithms designed (or watermarking) to detect inappropriate content. Moreover, we can develop and adhere to ethical guidelines and standards for the use of generative models. This might involve establishing principles for responsible AI development and deployment, including considerations for fairness, transparency, and accountability.

In summary, technology developed throughout this thesis both have exciting positive potential but may also lead to harm if used by bad actors. While we focus on improving technological aspects, governments and regulatory bodies should play a crucial role in making sure that technology is used responsibly and does not fall into the wrong hands.


[46] E Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. JASA, 1953.


[71] Hongliang Fei, Xu Li, Dingcheng Li, and Ping Li. End-to-end deep reinforcement learning based coreference resolution. In ACL, 2019.


[116] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. JAIR, 2013.


[137] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In ICML, 2021.


Bibliography


[183] Xiaofei Li, Laurent Girin, Sharon Gannot, and Radu Horaud. Multichannel speech separation and enhancement using the convolutive transfer function. TASLP, 2019.


