On the Alignment, Robustness, and Generalizability of Multimodal Learning

Jielin Qiu

CMU-CS-24-101

April, 2024

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA

Thesis Committee:
Christos Faloutsos (Co-chair)
Lei Li (Co-chair)
Yonatan Bisk
William Wang (University of California, Santa Barbara)

Submitted in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy

This research was sponsored in part by CMU CSD fellowships, the Defense Advanced Research Projects Agency (DARPA) ADAPTER program, Adobe Research Gift Funding, Allegheny Health Network, Mario Lemieux Center for Innovation and Research in EP, and Cleveland Clinic. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government, or any other entity.

Keywords: multimodal learning, semantic alignment, multimodal robustness, generalization, cross-domain alignment

Abstract

Multimodal intelligence, where AI systems can process and integrate information from multiple modalities, such as text, visual, audio, etc., has emerged as a key concept in today's data-driven era. This cross-modal approach finds diverse applications and transformative potential across industries. By fusing heterogeneous data streams, multimodal AI generates representations more akin to human-like intelligence than traditional unimodal techniques.

In this thesis, we aim to advance the field of multimodal intelligence by focusing on three crucial dimensions: multimodal alignment, robustness, and generalizability. By introducing new approaches and methods, we aim to improve the performance, robustness, and interpretability of multimodal models in practical applications. In this thesis, we address these critical questions: (1) How do we explore the inner semantic alignment between different types of data? How can the learned alignment help advance multimodal applications? (2) How robust are the multimodal models? How can we improve the models' robustness in real-world applications? (3) How do we generalize the knowledge of one learned domain to another unlearned domain?

This thesis makes contributions to all three technical challenges. We start with a contribution of learning cross-modal semantic alignment, where we explore establishing rich connections between language and image/video data, with a focus on the multimodal summarization task. By aligning the semantic content of language with visual elements, the resulting models can possess a more nuanced understanding of the underlying concepts. We delve into the application of Optimal Transport-based approaches to learn cross-domain alignment, enabling models to provide interpretable explanations of their multimodal reasoning process.

For the next contribution, we develop comprehensive evaluation metrics and methodologies to assess the robustness of multimodal models. By simulating distribution shifts and measuring the model's performance under different scenarios, we can gain a deeper understanding of the model's adaptability and identify potential vulnerabilities. We also adopt Optimal Transport to improve the model's robustness performance through data augmentation via Wasserstein Geodesic perturbation.

The third contribution revolves around the generalizability of multimodal systems, with an emphasis on the interactive domain and the healthcare domain. In the interactive domain, we develop new learning paradigms for learning executable robotic policy plans from visual observations by incorporating latent language encoding. We also use retrieval augmentation to make the vision-language models capable of recognizing and providing knowledgeable answers in real-world entity-centric VQA. In the healthcare domain, we bridge the gap by transferring the knowledge of LLMs to clinical ECG and EEG. In addition, we design retrieval systems that can automatically match the clinical healthcare signal to the most similar records in the database. This functionality can significantly aid in diagnosing diseases and reduce physicians' workload.

In essence, this thesis seeks to propel the field of multimodal AI forward by enhancing alignment, robustness, and generalizability, thus paving the way for more sophisticated and efficient multimodal AI systems.

Acknowledgments

These five years at CMU have been truly an amazing journey. I am incredibly grateful for the opportunity to express my profound appreciation to my advisors, Prof. Lei Li and Prof. Christos Faloutsos, for their unwavering support and invaluable mentorship throughout my educational and personal development journey. Their guidance during times of uncertainty and challenge was not just professional but also deeply personal, fostering a nurturing environment for my growth. Lei has been an exceptional advisor, teaching me the intricacies of critical thinking and research. His expertise and hands-on approach in guiding me through complex projects have been instrumental in honing my skills. Lei's ability to delve deep into impactful work and illuminate the path towards innovative solutions has had a lasting impact on my approach to research and problem-solving. Christos has been a pillar of support and wisdom. Whether I needed advice on academic matters or help in preparing for presentations, Christos was always there with his insightful guidance and unwavering support. His readiness to assist at any moment and his profound understanding of the academic landscape has greatly contributed to my confidence and preparedness in various academic endeavors. Their combined mentorship went beyond mere academic instruction; it was holistic guidance that encompassed all aspects of personal and professional development. Their consistent encouragement and faith in my potential provided me with the strength and resilience to navigate through challenging periods in my academic career. Their belief in my abilities served as a powerful catalyst, propelling me towards achieving my goals and surpassing expectations. Moreover, the kind and amiable personalities of Lei and Christos, combined with their meticulous and thoughtful approach to mentorship, have been a continuous source of inspiration. They have set an exemplary model for me to emulate in my future academic and professional engagements. The impact of their mentorship extends beyond the confines of academia; it has shaped my character and approach to life's challenges.

I am profoundly thankful for the valuable contributions of my thesis committee members, Prof. Yonatan Bisk and Prof. William Wang. The insightful feedback and comprehensive perspectives provided by Yonatan and William have not only enhanced the quality of my work but also encouraged me to explore more complex and broader intellectual areas. My association with Yonatan began when I was a teaching assistant in his course. He has imparted extensive knowledge and demonstrated effective ways of collaborating with others through his cheerful demeanor and genuine kindness. These qualities have consistently brought a positive atmosphere to our interactions, greatly enriching my learning experience. William has been exceptional in identifying the core significance of my work and offering invaluable advice for its enhancement. His proficiency in research and meticulous attention to detail have played a crucial role in my development as a researcher. Their support has not only aided me in achieving academic success but also in developing a deeper understanding and appreciation of the research process. Their influence extends beyond the confines of my thesis, leaving a lasting impact on my approach to academic and professional challenges.

I am immensely grateful for the mentorship and support I received from several academic mentors during my PhD studies. Each of them has played a unique and invaluable role in my development as a researcher and individual. I owe a special debt of gratitude to Prof. Bo Li from the University of Chicago. Under Bo's guidance for two years, I have received consistent, insightful advice and suggestions that significantly improved my work. Bo's vast knowledge across various fields and ability to pinpoint the most critical aspects of a problem have been incredibly influential in my academic growth. I am also deeply appreciative of Prof. Doug Weber's generosity. His funding support for two years and his ongoing encouragement were pivotal in facilitating my research endeavors. Prof. Daniel Fried deserves special mention for his detailed and constructive feedback on my work. Our discussions were always enlightening, and I learned a great deal from his guidance on crafting high-quality research papers. Daniel's meticulous approach to academic discourse has been a great source of learning for me. My gratitude extends to Prof. Jun-Yan Zhu as well, with whom I formed both a mentorship and a friendship. Since meeting Jun-Yan as a teaching assistant in his course, I have gained immense knowledge. His advice, covering both academic and personal aspects, especially during uncertain times in my life, has been incredibly valuable. I am also thankful for the mentorship of Dr. Emerson Liu from Allegheny Health Network and Dr. Michael Rosenberg from the University of Colorado Cardiovascular Institute. Their teachings in the clinical domain laid the foundation for my work in healthcare, providing me with the essential knowledge needed to bridge the gap between technical and clinical aspects. The opportunity to learn from Prof. XuanLong Nguyen has been another highlight of my PhD journey. XuanLong exemplified what it means to be a dedicated and decent researcher, setting a standard for me to aspire to in my academic career. I also greatly appreciate Prof. Tianqi Chen and Prof. Nancy Pollard for helping me with my CSD Speaking Skills. Lastly, my heartfelt thanks go to Prof. Fei Fang. Her support and guidance during challenging times have been extraordinary. Fei's positive attitude and bright smile have a way of making complex problems seem more manageable, providing me with a perspective that often led to better solutions than I could have imagined. Her influence has been a beacon of optimism in my academic journey. I am also fortunate to learn from senior researchers, including Dr. Xiuming Zhang from MIT, Dr. Alex Smola from Boson AI, Dr. Tristan Naumann, Dr. Ozan Oktay from Microsoft, Prof. Weina Wang, Prof. Ding Zhao, Prof. Jeff Schneider, Prof. Leila Wehbe, Prof. Zico Kolter, Prof. Changliu Liu, Prof. David Held, Prof. Leman Akoglu, Prof. Steven Brookes from CMU.

I am profoundly grateful for the mentorship and support I received from numerous esteemed mentors during my internships, each of whom has significantly contributed to my professional and personal growth. I extend my sincere thanks to Dr. Yi Zhu, who has been an outstanding mentor since my time interning at AWS. Yi's insights and guidance were crucial in nurturing my development during crucial stages of my academic journey. Alongside Dr. Xingjian Shi, also at AWS, I was fortunate to gain extensive knowledge and research skills and experience their warm-hearted approach to mentorship. My gratitude also goes to Dr. Hailin Jin, whose exceptional support over the past three years has been instrumental in my development. As my manager during my internship at Adobe, Hailin offered me invaluable guidance and the opportunity to delve into a new research topic that later contributed to a patent. His availability and willingness to assist even after my internship are deeply appreciated. The internship at Meta was another enriching experience, where I learned immensely from Dr. Luna Dong, Dr. Shane Moon, Dr. Zhaojiang Lin, and Dr. Andrea Madotto. The team at Meta provided an environment that exceeded my expectations, offering a unique and valuable learning experience. Across my five internships, I have been privileged to work with and learn from a group of exceptional mentors, including Dr. Hailin Jin, Dr. Zhaowen Wang, Dr. Trung Bui, Dr. Franck Dernoncourt from Adobe; Dr. Yi Zhu, Dr. Xingjian Shi, Dr. Zhiqiang Tang, Dr. Mu Li from Amazon Web Service; Dr. Jianfeng Wang, Linjie Li, Dr. Zhengyuan Yang, Dr. Lijuan Wang from Microsoft; Dr. Shane Moon, Dr. Andrea Madotto, Dr. Zhaojiang Lin, Dr. Luna Dong, Dr. Ethan Xu, Dr. Paul Crook, Babak Damavandi from Meta; and Shangbang Long from Google. The opportunity to work alongside and learn from these brilliant individuals has been an absolute pleasure and an enriching experience.

I am filled with heartfelt gratitude for my collaborators and friends, whose presence has greatly enriched both my academic journey and personal life. Their insightful discussions, unwavering support, and shared moments of joy have been invaluable. Special thanks to my friend Jiacheng Zhu, whose reliability and inspiring work ethic have continually motivated me. To Mengdi Xu, with whom I have shared not only life experiences but also cultivated research ideas and sought advice, my unending gratitude. I am grateful for the climbing adventures and great moments shared with Zuxin Liu and Rui Chen. These experiences have added a unique and enjoyable dimension to my life. My profound appreciation equally goes to close friends Peide Huang, Yiwen Dong, Hanjiang Hu, Yuyou Zhang, Wenhao Ding, Miao Li, Haohong Lin, Laixi Shi, Yaru Niu, William Han, Shiqi Liu, Zhepeng Cen, Yihang Yao, Changyi Lin, Diana Gomez – with whom I have forged a bond akin to family, supporting each other through thick and thin. I also extend thanks to my labmates Zhenqiao Song, Danqing Wang, Xuandong Zhao, Siqi Ouyang, Kexun Zhang, and Wenda Xu for their support and camaraderie. Additionally, I cherish the friendships formed within the CSD community, including with Yi Zhou, Mingjie Sun, Shuqi Dai, Chen Dan, Dorian Chan, Long Pham, Minji Yoon, and my office mates Lucio Dery, Asher Trockman, Praneeth Kacham, and Suhas Jayaram Subramanya. I am equally thankful for the companionship of friends like Scarlett Zhang, Thomas Ning, Weiye Zhao, Tian Lu, Chengyuan Zhang, Jingqi Zhang, Zilin Si, Yeeho Song, Helen Jiang. Furthermore, I have been fortunate to have close friends and collaborators outside CMU, such as Cindy Wu from Princeton, Bo He, Jun Wang from UMD, Zhuolin Yang from UIUC, Zijie Huang from UCLA, Aritra Guha from the University of Michigan, Makiya Nakashima, Jaehyun Lee, Debbie Kwon, David Chen from Cleveland Clinic, and Florian Wenze from Amazon. The shared experiences, laughter, and support from all these wonderful individuals have made my journey not only successful but also joyous and fulfilling. Their companionship has been a source of immense joy and a key factor in my personal and professional growth.

I want to express my deep gratitude to the Computer Science Department for their support and encouragement. A heartfelt thank you to Prof. Karl Crary and Prof. Srinivasan Seshan for their steadfast belief in me and their generous assistance, especially during challenging periods. Their unwavering support has been a fundamental pillar in my academic journey. I am also profoundly thankful to Deb Cavlovich and Matthew Stewart for their invaluable and continuous support, which has greatly simplified many complex tasks. Their assistance has been crucial in navigating the intricacies of academic processes. Additionally, my appreciation goes to Nick Hernandez, Jenn Landefeld, Amanda Hornick, and Tracy Farbacherfor their timely support and commitment. Their efforts have played a significant role in fostering a smooth and supportive academic environment, making my journey both successful and enjoyable.

In closing, my most profound and heartfelt thanks are reserved for my entire family, whose constant love and support have been the foundation of everything I have accomplished. To my parents, Guoming and Hua, my gratitude is immense and truly without limits. Their endless love, unwavering support, and consistent encouragement have been the guiding forces in my life. Their steadfast belief in me has illuminated every step of my journey. To my beloved wife, Zhenni, your unwavering encouragement and profound faith in me have served as the cornerstone of my strength. Your presence in my life has not merely illuminated my journey but has consistently been a wellspring of inspiration and solace, enriching every step I take. Special thanks to my cat, who has been my unwavering buddy, offering comfort and companionship through every milestone and tearful moment. This journey, brimming with triumphs and challenges, would have been unimaginable without your unwavering love and support. This achievement reflects not only my efforts but equally your sacrifices and steadfast backing. I am deeply grateful for your constant support and the numerous sacrifices you have made. To all of you, my gratitude knows no bounds.

To Guoming, Hua, and Zhenni

Contents

List of Figures

List of Tables

List of Algorithms

List of Acronyms

Chapter 1

Introduction

The world is fulfilled with various signals such as images, videos, audio, text, sensor signals, and so on. Humans naturally perceive the world through multiple senses, such as sight, sound, touch, and more. This multi-modal approach is a fundamental aspect of human cognition and has become a cornerstone in the development of artificial intelligence (AI). In the current era, where technology is increasingly driven by vast amounts of data, the concept of multimodal intelligence has gained prominence. This paradigm combines various forms of data, such as language, visual information, physiological signals, and others, to create more comprehensive and nuanced AI systems.

Despite the wealth of data available, there are currently no comprehensive tools for analyzing this data and leveraging the patterns within it. Identifying patterns across different modalities is essential for utilizing them to address real-world, domain-specific challenges. For motivation examples: (1) Summarization and Recommendation: The automatic generation of summaries for multimedia news or providing introductions to online videos can significantly enhance the performance of search engine and recommendation systems for online content. For instance, more than 500 hours of video are uploaded to YouTube every minute, most without summaries. The absence of accurate summaries makes it challenging to develop search engines and recommendation systems that efficiently help users find the content they desire. (2) Content Creation: Text-to-image generation models often produce incorrect, unclear, or biased content. Developing more robust models for real-world applications is an urgent issue. (3) Household Robots: The market for household robots, aimed at assisting people with disabilities in daily tasks, was valued at USD 10.3 billion in 2023. These robots aim to offer substantial and efficient support for routine tasks through various interaction methods, including visual perception, verbal instructions, and speech dialogue, making the utilization of multimodal information to enhance the robot's performance and reliability a complex challenge. (4) Healthcare Applications: For example, the diagnosis of chest pain in emergency departments (ED) alone currently incurs an estimated cost of 10 to 12 billion per year in the US. Developing a solution that provides cost-efficient patient care using multimodal healthcare data could be highly beneficial for society.

This thesis is motivated by these applications. The problems studied in this thesis are abstracted from the common challenges across these applications. In the following, we will first present a few motivating application. We will describe the problems and general approaches to the challenges.

1.1 Motivation and Challenges

Multimodal Learning has advanced quickly in recent years with numerous applications in different fields, i.e., multimedia image/video and language understanding [99, 179, 389, 580], embodied learning [38, 178, 192, 294], healthcare and psychology [146, 267], and many more.

Multimodal learning is essential in real-world applications as it empowers systems to process and comprehend information from a variety of sources. In these scenarios, it is very important to understand the patterns in the data such as alignment, robustness, and generalizability. Our goal is to develop algorithms and build datasets for learning from multiple modalities, and we list here a few motivating applications.

Multimedia search engine and recommendation systems The digital world is overflowing with multimedia content, such as videos. For instance, with more than 500 hours of video uploaded to YouTube every minute, many without detailed annotations, creating multimedia search engines and recommendation systems is a challenging task. Traditional search engines rely on titles, text descriptions, or video tags, but for more precise search results, additional content like summaries is needed. However, the number of videos with summaries is small compared to the volume of new uploads, making it crucial to find ways to generate summaries for videos to enhance search engine performance and better recommendations for the users.

One promising solution is Multimodal Summarization with Multimodal Output (MSMO), which has gained traction in recent years [52, 172, 210, 309, 576]. MSMO aims to automatically generate keyframes and key textual summaries for media news or online videos. These applications can deliver concise summaries of multimedia content, which can significantly improve the development of search engines and recommendation systems, helping users find the content they are looking for more effectively.

Figure 1.1: Multimedia summarization: providing summaries can significantly lead to better search engines and recommendation systems.

Figure 1.1 illustrates a potential solution for addressing the randomness of videos on the internet, which can aid in the development of improved search engines and recommendation systems. In this context, we are particularly interested in addressing the following critical issues:

- How do we explore the inner semantic alignment between different domains?

- How can the learned alignment help advance multimodal applications, such as providing better summarization results?

Content Creation As the field of text-to-image generation models [317, 389] evolves, these technologies are increasingly being utilized to assist in the creation of creative content. They serve a wide range of users, from professional designers to individuals without specific domain expertise, effectively enhancing the efficiency and creativity of content production. Despite their growing popularity, these models encounter several challenges. One of the main issues is their vulnerability to attacks, which can lead to the generation of incorrect, unclear, or biased content. This highlights the pressing need for the development of more robust models that can withstand such vulnerabilities and perform reliably in real-world applications. Addressing this concern is essential for ensuring the utility of text-to-image generation technologies in various creative fields.

Figure 1.2: Content creation: text-to-image models are easy to attack and generate incorrect results. For example, they might generate an image of a dog that is not white or a scene without grass or trees, despite these elements being specified in the input description but with a few common perturbations. (“Keyboard" simulates the mistakes made while using a keyboard.)

Figure 1.2 illustrates a text-to-image generation example, where applying keyboard typos, OCR errors, or synonym replacements to the original sentence, can lead to generated images containing incomplete visual information. In this context, we are interested in exploring:

- How robust are the multimodal models in the presence of noises??
- How can we improve the models' robustness in real-world applications?

Household Robot In 2023, the industry focusing on household robots designed to aid individuals with disabilities in their daily activities was estimated to be worth USD 10.3 billion. These robots aim to offer substantial and efficient support for routine tasks through various interaction methods, including visual perception, verbal instructions, and speech dialogue, making the utilization of multimodal information to enhance the robot's performance and reliability a complex challenge.

Large Language Models (LLMs) have recently made significant advancements in supporting robotic learning for intricate domestic tasks such as complex household management. Nonetheless, the efficacy of these pre-trained LLMs depends greatly on templated text data tailored to specific domains, a requirement that may not be practical for real-world robotic learning scenarios that involve image-based observations. Furthermore, current LLMs that process textual information are not designed to adapt through non-expert interactions with environments. Consequently, the pressing question arises: How can we leverage multimodal data to develop better household robots that can more effectively assist humans in the complex environment?

Figure 1.3: Household Robot: the market is increasing with few practical solutions [3].

Healthcare applications Providing high-quality and efficient patient treatment is a longstanding problem. For example, in the current practice of cardiovascular disease, patients presenting with chest pain to the emergency department (ED) constitute a diagnostic and logistic challenge as chest pain can be caused by an extensive variety of disorders [14]. Diagnostic tests and decision algorithms play a critical role in speeding up the appropriate triage of chest pain patients in the ED, and preventing unnecessary hospitalization of patients with non-critical disorders. In current practice, about half of the patients presenting with chest pain can be discharged from the ED, and only 5.5 % of all ED visits lead to serious diagnosis [175]. However, the diagnosis of chest pain in the ED now incurs an estimated cost of 10 to 12 billion per year in the U.S., representing a significant financial burden for both patients and society. The pressing issue, therefore, is how to provide cost-efficient patient care using multimodal healthcare data.

Figure 1.4: Healthcare applications: the healthcare treatment cost is a financial burden but still increasing.

Since the data available in the robotics and healthcare domains are less abundant compared to the vast amounts of multimedia image/video or text data, there are inherent challenges, even though healthcare data, as illustrated in Figure ??, comes in various types. However, the volume of

Figure 1.5: Healthcare applications: different types of healthcare data, from [372].

data within each type is still signi cantly lower than that of image-text data. This leads to typical problems including:

- How do we generalize the knowledge of one learned domain to another unlearned domain?
- How to generalize from data-rich domain to data-scarce domain?

1.2 Thesis Overview

Figure 1.6: Thesis contributions to multimodal alignment, robustness, and generalizability.

Given this context, several critical questions need to be addressed to improve the alignment, robustness, and generalizability of multimodal learning:

1. Exploring Inner Semantic Alignment How do we explore the inner semantic alignment between different domains? How can the learned alignment help advance multimodal ap-

plications? Understanding the relationships between different modalities and how they can complement each other is crucial for developing more effective multimodal systems.

2. Robustness of Multimodal Models How robust are the multimodal models? How can we improve the models' robustness in real-world applications? Ensuring that multimodal models can handle diverse and potentially noisy inputs is essential for their reliability and effectiveness in practical applications.
3. Generalization Across Domains How do we generalize the knowledge of one learned domain to another unlearned domain? Developing methods that allow models to transfer learned knowledge to new, unseen domains is key to creating more versatile and adaptable multimodal systems.

In this thesis, we aim to answer each of these three questions, aiming to enhance the performance, robustness, and interpretability of multimodal models in real-world scenarios, ultimately contributing to the advancement of multimodal intelligence.

Chapter 2: multimodal semantic alignment This chapter focuses on achieving effective multimodal semantic alignment, facilitating seamless connections between language and visual modalities. To accomplish this objective, the following sub-objectives are pursued:

- Multimodal alignment: We explore establishing rich connections between language and image/video data. By aligning the semantic content of language with visual elements, the resulting models can possess a more nuanced understanding of the underlying concepts. In [361], we propose a Semantics-Consistent Cross-domain Summarization (SCCS) model, which leverages optimal transport alignment combined with visual and textual segmentation to achieve multimodal summarization. In [347], we explore the alignment between visual and language domains speci cally for the task of temporally segmenting long Livestream videos, which can establish the basis for Livestream video understanding tasks and can be extended to many real-world applications.
- Interpretability: We delve into the application of Optimal Transport-based approaches to learn cross-domain alignment, enabling models to provide interpretable explanations of their multimodal reasoning processes [361, 363]. The optimal transport coupling can reveal the underlying similarity and structure, which further helps to explain the correspondence between the text and image data.
- New datasets: We propose a new dataset, MMSum [357], to solve the problems within existing datasets, such as insuf cient maintenance, data inaccessibility, limited size, etc. MMSum is speci cally designed to cater to a wide range of tasks, with a particular emphasis on MSMO, with diverse categorization.

Chapter 3: multimodal robustness In Chapter 3, we aim to address the challenge of multimodal robustness, particularly under perturbations. As real-world scenarios often involve variations in data distributions, it is crucial to ensure that multimodal models can maintain their performance across diverse environments. To tackle this challenge, the research will focus on the following areas:

- Robustness evaluation: In [363], we build a comprehensive evaluation benchmark speci cally designed to assess the robustness of multimodal models. By simulating various distribution shifts and measuring the model's performance across different scenarios, we aim to gain a deeper insight into the model's adaptability and vulnerabilities. Understanding how multimodal models perform under diverse conditions is crucial for developing more reliable and robust systems that can effectively handle the complexities of real-world data.

Chapter 4: multimodal generalizability capabilities in interactive environments In Chapter 4, we aim to explore the multimodal generalizability capabilities in interactive environments, particularly in the domains of language grounding. The following sub-objectives are pursued:

- Language grounding in robot learning: In [350], we introduce a novel learning paradigm that generates robots' executable actions in the form of text, derived solely from visual observations. Our proposed paradigm stands apart from previous works, which utilized either language instructions or a combination of language and visual data as inputs.
- Retrieval-augmented generation (RAG): In [353], we work on a novel task for entity-centric VQA to assess the proficiency of models in accurately identifying and generating responses that exhibit a deep comprehension of these identified entities. We propose a retrieval-augmented multimodal LLM, devised as a baseline model capable of undertaking the SnapNTell [353] task, which is scalable, effective, and explainable.

Chapter 5: cross-modal applications in healthcare Finally, in Chapter 5, we explore whether the learned knowledge can be transferred to the clinical domain. We aim to explore the cross-modal applications in healthcare.

- ECG-to-text generation: In [349], we aim to bridge the gap by transferring the knowledge of LLMs to clinical Electrocardiography (ECG) for textual diagnosis report generation and zero-shot disease detection. Our approach is able to generate high-quality cardiac diagnosis reports and also achieves competitive zero-shot classification performance even compared with supervised baselines, which proves the feasibility of transferring knowledge from LLMs to the cardiac domain.
- Connectivity between human language and brain signals: In [146], we explore the relationship and dependency between EEG and human language to reveal the inner connection. Our findings on word-level and sentence-level EEG-language alignment show the influence of different language semantics as well as EEG frequency features.
- Clinical retrieval system: In [351], we design a retrieval system that can automatically match the input Cardiovascular Magnetic Resonance (CMR) Imaging to the most similar records in the database. This functionality can significantly aid in diagnosing diseases and reduce physicians' workload, which can provide patients with better treatment.

1.3 Summary of Contributions

Multimodal intelligence, where AI systems exhibit intelligent behaviors by leveraging data from multiple modalities (text, visual, audio, etc.), has emerged as a key concept in today's data-driven era. This cross-modal approach finds diverse applications and transformative potential across industries. By fusing heterogeneous data streams, multimodal AI generates representations more akin to human-like intelligence than traditional unimodal techniques. In this thesis, we aim to propel the field of multimodal AI forward by enhancing alignment, robustness, and generalizability, thus paving the way for more sophisticated and efficient multimodal AI systems.

Algorithms Our work has focused on:

- Multimodal alignment [347, 357, 361]: We explore establishing rich semantic connections between language and image/video data, with a focus on Multimodal Summarization with Multimodal Output (MSMO) task. By aligning the semantic content of language with visual elements, the resulting models can possess a more nuanced understanding of the underlying concepts.
- Interpretability [361, 363]: We delve into the application of Optimal Transport-based approaches to learn cross-domain alignment, enabling models to provide interpretable explanations of their multimodal reasoning process.
- Language grounding in robot learning [355]: This research aims to develop techniques for learning executable plans from visual observations by incorporating latent language encoding. Models are trained to understand and interpret visual cues while leveraging the rich semantic information encoded in language.
- Retrieval-augmented Multimodal LLM [353]: We develop a retrieval-augmented Multimodal LLM model, which is capable of recognizing and providing knowledgeable answers in real-world entity-centric Visual Question Answering (VQA).
- ECG-to-text generation [349]: We bridge the gap by transferring the knowledge of LLMs to clinical ECG for diagnosis report generation and zero-shot disease detection.
- Connection between human language and brain signals [146]: We explore the relationship and dependency between EEG and human language to reveal the inner connection.
- Clinical retrieval system for Cardiovascular Magnetic Resonance (CMR) Imaging [351]: We design a retrieval system that can automatically match the input signal to the most similar records in the database. This functionality can signi cantly aid in diagnosing diseases and reduce physicians' workload.

Datasets and Benchmark Our work has focused on:

- Robustness evaluation benchmark of multimodal models [363]: We develop comprehensive evaluation metrics and methodologies to assess the robustness of multimodal models. By simulating distribution shifts and measuring the model's performance under different scenarios, we can gain a deeper understanding of the model's adaptability and identify potential vulnerabilities.
- New MSMO dataset [357]: We propose a new dataset named MMSum to solve the problems within existing MSMO datasets, such as insuf cient maintenance, data inaccessibility, limited size, and categorization, etc., spanning 17 principal categories and 170 subcategories.
- New Livestream video dataset [347]: We introduce a new large dataset of Livestream videos, which contains 11,285 Livestream videos with a total duration of 15,038.4 hours.
- New CMR dataset [351]: The existing work falls short in providing a large CMR dataset, we take the initiative to gather a comprehensive dataset consisting of 13,787 studies derived from actual clinical cases.
- New entity-centric VQA dataset [353]: We have developed the SnapNTell dataset, distinct from traditional VQA datasets as (1) It encompasses a wide range of categorized entities,

each represented by images and explicitly named in the answers; (2) It features QA pairs that require extensive knowledge for accurate responses. The dataset is organized into 22 major categories, containing 7,568 unique entities in total.

Applications The works in this thesis are listed in Table 1.1 based on the topics.

Table 1.1: Thesis topics.

	Model/Algorithm	Dataset/Benchmark	Application Multimedia	Robotics	Healthcare	Venue
Alignment	SCCS LiveSeg MMSum	LiveSeg MMSum Entity6K	SCCS LiveSeg MMSum Entity6K			ACL Findings 2023 [361] WACV 2023 [347] CVPR 2024 [357] Under Review [348]
Robustness	MMRobustness Cardiac-MT Interp-OT GeoECG	MMRobustness	Interp-OT		Cardiac-MT GeoECG	DMLR 2024 [363] ICASSP 2023 [362] ICML 2023 [574] PMLR MLHC 2022 [575]
Generalization	SUM+APM ECG-LLM MTAM CMRformer ECG-Encoding SnapNTell	CMRformer SnapNTell	SnapNTell	SUM+APM	ECG-LLM MTAM CMRformer ECG-Encoding	NAACL 2024 [355] EACL Findings 2023 [349] EMNLP Findings 2023 [146] ICML 2023 Workshop [351] ML4H 2023 [358] Under review [353]

The work presented in this thesis has already had a signi cant impact, which we categorize into 1) Application impact, 2) Academic impact, and 3) Impact through code release.

Application Impact

- [347] has been patented by Adobe for production on Behance Livestream [197].
- [351] has been implemented by Cleveland Clinic for clinical trials.

Academic Impact The work presented here has been previously published in top-tier outlets in different venues, as in Table 12.1, especially:

- [363] was accepted as the very rst paper of the Journal of Data-centric Machine Learning Research (DMLR).
- [357] was accepted as Poster (Highlight) in CVPR 2024, which is Top 11.9% among all accepted papers.
- [355] was accepted as a spotlight for the ICML 2023 Workshop on Interactive Learning with Implicit Human Feedback.

Impact Through Code Release To enable the reproducibility and extendibility of our work, we have publicly released the source code for the algorithms presented in this thesis. As of January 28, 2024:

- [363] has 30 stars, 18 clones, 173 viewers, and 17 citations (as of March 24, 2024).
- [357] has 24 stars, 16 clones, and 424 viewers (as of March 24, 2024).
- [267] has 141 citations.

Other Contributions Here we list some other contributions that happened during the graduate study but are not introduced in the thesis.

- Data augmentation: Data augmentation techniques can improve the models' robustness. In [362, 458, 575], we synthesize diverse examples encompassing a wide range of data distributions; models can learn to generalize better and exhibit improved performance in novel scenarios. We propose a physiologically-inspired data augmentation method to improve the performance, generalization, and robustness of the ECG prediction model, where the new data augmentation method was proposed from a probability perspective. We perturb the data distribution towards other classes along the geodesic in a Wasserstein space. Also, the ground metric of this Wasserstein space is computed via a set of physiological features so that the perturbation lies on a manifold that exploits the physiological properties.
- ECG-encoding: In [358], we encode ECG as images and adopted a vision-language learning paradigm to jointly learn vision-language alignment between encoded ECG images and ECG diagnosis reports. Encoding ECG into images can result in an efficient ECG retrieval system, which can be highly practical and useful in clinical applications.
- Geometry-aware representations with Wasserstein distance: In [574], we learn diverse representations based on the Wasserstein distance. In particular, by choosing the distance metric in the primal definition of the Wasserstein distance to reflect our prior knowledge about the data, we can formulate the corresponding dual problem and learn diverse representations, maximizing the pairwise Wasserstein distances under certain model smoothness constraints.

Thesis Statement Multimodal intelligence refers to the ability of a system to integrate and process information from multiple sensory modalities, such as visual, text, tactile inputs, etc., to achieve complex understanding and perform tasks. This type of intelligence is crucial in environments where diverse types of data are present, enabling effective interaction, decision-making, and problem-solving by leveraging the complementary strengths of each modality.

This thesis demonstrates that by enhancing the modeling of patterns across different modalities, we can achieve a more effective utilization of the unique modality equivalence learned through abstract multimodal representations. This improved modeling can lead to advancements in cross-modal applications, increasing the robustness of multimodal models under distribution shifts and enhancing their generalization abilities. Consequently, this thesis aims to advance the field of multimodal AI by focusing on the enhancement of alignment, robustness, and generalizability, ultimately leading to the development of more sophisticated and efficient multimodal AI systems.

However, a notable challenge in this domain is the opaque nature of these complex models. The internal logic behind their alignment across different modalities is often not transparent, posing difficulties in understanding and interpreting their behavior. Future research in this area could delve into the interpretability of multimodal AI systems, exploring methods to elucidate the alignment logic across different modalities and how it can be leveraged more efficiently.

Chapter 2

Background and Preliminary

This chapter reviews the current literature of multimodal learning related work. We review general methodologies and models in the literature. More specific ones will be introduced later in each chapter.

2.1 Background

Multimodal Learning has advanced quickly in recent years with appealing applications in different fields, i.e., embodied learning [38, 178, 192, 294], multimedia image/video and language understanding [99, 179, 389, 580], and psychology [146, 267]. Thanks to the larger datasets [334, 367, 406, 407, 539] and larger transformer models [43, 59, 68, 253, 546], many powerful multimodal image-text models have been developed and shown great capability. However, unlike unimodal models, the robustness study of multimodal models under distribution shift has rarely been explored.

Multimodal Alignment Aligning representations from different modalities is important in multimodal learning. Exploring the explicit relationship across vision and language has drawn significant attention [481]. [461, 514, 538] adopted attention mechanism, [90] composed pairwise joint representation, [57, 502, 547] learned fine-grained or hierarchical alignment, [231, 503] decomposed the inputs into sub-tokens, [470, 530] adopted graph attention for reasoning, and [141, 367, 465, 522] applied contrastive learning algorithms.

Multimodal Summarization explored multiple modalities for summary generation by learning the alignment [114, 324, 494, 541] learned the relevance or mapping in the latent space between different modalities. In addition to only generating visual summaries, [19, 236, 576] generated textual summaries by taking audio, transcripts, or documents as input along with videos or images, using seq2seq model [445] or attention mechanism [24]. The methods above explored using multiple modalities' information to generate a single modality output, either textual or visual summary. Recent trends on the MSMO task have also drawn much attention [114, 115, 154, 295, 360, 361, 452, 552, 557, 576]. Specifically, [452] summarized a video and text document into a cover frame and a one-sentence summary. The most significant difference between multimodal summarization and MSMO lies in the inclusion of multiple modalities in the output.

Multimodal LLMs Expanding text-only LLMs to interpret visual information typically involves integrating a visual encoder with a frozen LLM, using extensive image captioning datasets for alignment [68, 219, 504]. This integration can be accomplished through methods such as adapter-based tuning [8], which fine-tunes a small portion of the model to process visual inputs, or prefix tuning [463], where trained prefixed vectors are inputted to guide the frozen LLM towards contextually relevant text outputs based on the visual data. These techniques allow LLMs to maintain their linguistic prowess while gaining visual understanding without full model retraining [534].

2.2 Preliminary

Several fundamental methodologies have been established in the literature. In this section, we present their preliminaries, which will be utilized in subsequent chapters.

Optimal Transport (OT) Basis OT is the problem of transporting mass between two discrete distributions supported on latent feature space $\mathcal{X}$. Let $\mu = \{x_i, \mu_i\}_{i=1}^{n}$ and $\nu = \{y_j, \nu_j\}_{j=1}^{m}$ be the discrete distributions of interest, where $x_i, y_j \in \mathcal{X}$ denotes the spatial locations and μ_i, ν_j, respectively, denoting the non-negative masses. Without loss of generality, we assume $\sum_i \mu_i = \sum_j \nu_j = 1$. $\pi \in \mathbb{R}_{+}^{n \times m}$ is a valid transport plan if its row and column marginals match μ and ν, respectively, which is $\sum_i \pi_{ij} = \nu_j$ and $\sum_j \pi_{ij} = \mu_i$. Intuitively, π transports π_{ij} units of mass at location x_i to new location y_j. Such transport plans are not unique, and one often seeks a solution $\pi^* \in \Pi(\mu, \nu)$ that is most preferable in other ways, where $\Pi(\mu, \nu)$ denotes the set of all viable transport plans. OT finds a solution that is most cost-effective w.r.t. cost function $C(x, y)$:

$$D(\mu, \nu) = \sum_{ij} \pi^*_{ij} C\left(x_i, y_j\right) = \inf_{\pi \in \Pi(\mu, \nu)} \sum_{ij} \pi_{ij} C\left(x_i, y_j\right) \tag{2.1}$$

where $D(\mu, \nu)$ is known as OT distance. $D(\mu, \nu)$ minimizes the transport cost from μ to ν w.r.t. $C(x, y)$. When $C(x, y)$ defines a distance metric on $\mathcal{X}$, and $D(\mu, \nu)$ induces a distance metric on the space of probability distributions supported on $\mathcal{X}$, it becomes the Wasserstein Distance (WD).

Wasserstein Distance As above, Wasserstein distance (WD) is introduced in OT, which is a natural type of divergence for registration problems as it accounts for the underlying geometry of the space, and has been used for multimodal data matching and alignment tasks [54, 81, 230, 360, 540, 575]. In Euclidean settings, OT introduces WD $\mathcal{W}(\mu, \nu)$, which measures the minimum effort required to "displace" points across measures μ and ν, where μ and ν are values observed in the empirical distribution. In our setting, we compute the temporal-pairwise Wasserstein Distance on EEG features and language features, which are $(\mu, \nu) = (V_e, V_t)$. For simplicity without loss of generality, assume $\mu \in \mathrm{P}(\mathbb{X})$ and $\nu \in \mathrm{P}(\mathbb{Y})$ denote the two discrete distributions, formulated as $\mu = \sum_{i=1}^{n} u_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} v_j \delta_{y_j}$, with δ_x as the Dirac function centered on x. $\Pi(\mu, \nu)$ denotes all the joint distributions $\gamma(x, y)$, with marginals $\mu(x)$ and $\nu(y)$. The weight vectors $u = \{u_i\}_{i=1}^{n} \in \Delta_n$ and $v = \{v_i\}_{i=1}^{m} \in \Delta_m$ belong to the n- and m-dimensional simplex, respectively. The WD between the two discrete distributions μ and ν is defined as:

$$\mathcal{WD}(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu)} \mathbb{E}_{(x, y) \sim \gamma}[c(x, y)] = \min_{T \in \Pi(u, v)} \sum_{i=1}^{n} \sum_{j=1}^{m} T_{ij} \cdot c\left(x_i, y_j\right) \tag{2.2}$$

where $\Pi(u, v) = \{T \in \mathbb{R}_+^{n \times m} \mid T 1_m = u, T^\top 1_n = v\}$, 1_n denotes an m dimensional all-one vector, and $c(x_i, y_j)$ is the cost function evaluating the distance between x_i and y_j.

Canonical Correlation Analysis Canonical Correlation Analysis (CCA) is a method for exploring the relationships between two multivariate sets of variables. It learns the linear transformation of two vectors to maximize the correlation between them, which is used in many multimodal problems [16, 129, 352]. In this chapter, we apply CCA to capture the cross-domain relationship. Let low-level transformed EEG features be V_e and low-level language features be L_t. We assume $(V_e, V_t) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ has covariances $(\Sigma_{11}, \Sigma_{22})$ and cross-covariance Σ_{12}. CCA finds pairs of linear projections of the two views, $(w_1' V_e, w_2' V_t)$ that are maximally correlated:

$$(w_1, w_2) = \operatorname*{argmax}_{w_1, w_2} \operatorname{corr}(w_1' V_e, w_2' V_t) = \operatorname*{argmax}_{w_1, w_2} \frac{w_1' \Sigma_{12} w_2}{\sqrt{w_1' \Sigma_{11} w_1 w_2' \Sigma_{22} w_2}} \tag{2.3}$$

Transformer Architecture The transformer is based on the attention mechanism [468]. The original transformer model is composed of an encoder and a decoder. The encoder maps an input sequence into a latent representation, and the decoder uses the representation with other inputs to generate a target sequence.

First, we feed out the input into an embedding layer, which is a learned vector representation of the input feature, by mapping the features to a vector with continuous values. Then we inject positional information into the embeddings by:

$$PE_{(pos, 2i)} = \sin\left(pos / 10000^{2i/d_{model}}\right), \; PE_{(pos, 2i+1)} = \cos\left(pos / 10000^{2i/d_{model}}\right) \tag{2.4}$$

The attention model contains two sub-modules, a multi-headed attention model and a fully connected network. The multi-headed attention computes the attention weights for the input and produces an output vector with encoded information on how each feature should attend to all other features in the sequence. There are residual connections around each of the two sub-layers followed by a layer normalization, where the residual connection means adding the multi-headed attention output vector to the original positional input embedding, which helps the network train by allowing gradients to flow through the networks directly.

Multi-headed attention applies a self-attention mechanism, where the input goes into three distinct fully connected layers to create the query, key, and value vectors. The output of the residual connection goes through a layer normalization.

$$\begin{aligned} \operatorname{MultiHead}(Q, K, V) &= \operatorname{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h) W^O \\ \text{where } \mathrm{head}_i &= \operatorname{Attention}\left(Q W_i^Q, K W_i^K, V W_i^V\right) \end{aligned} \tag{2.5}$$

The attention model contains N same layers, and each layer contains two sub-layers, which are a multi-head self-attention model and a fully connected feed-forward network. Residual connection and normalization are added in each sub-layer. So the output of the sub-layer can be expressed as: $\mathrm{Output} = \operatorname{LayerNorm}(x + (\operatorname{SubLayer}(x)))$ For the Multi-head self-attention module, the attention can be expressed as $\mathrm{attention} = \operatorname{Attention}(Q, K, V)$, where multi-head attention uses h different linear transformations to project query, key, and value, which are Q, K, and V, respectively, and finally concatenate different attention results:

$$\operatorname{MultiHead}(Q, K, V) = \operatorname{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h) W^O \tag{2.6}$$

$$\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.7}$$

where the projections are parameter matrices:

$$W_i^Q \in \mathbb{R}^{d_{model}\times d_k}, \quad W_i^K \in \mathbb{R}^{d_{model}\times d_k}, W_i^V \in \mathbb{R}^{d_{model}\times d_v}, \quad W_i^O \in \mathbb{R}^{hd_v \times d_{model}} \tag{2.8}$$

where the computation of attention adopted scaled dot-product:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.9}$$

Vision Transformer (ViT) A Vision Transformer (ViT) [92] is a type of transformer specifically developed for computer vision tasks. Unlike text-based transformers that split text into tokens, a ViT divides an input image into several patches. Each patch is then converted into a vector and reduced to a smaller dimension through matrix multiplication. These vector embeddings are processed by a transformer encoder, similar to how token embeddings are handled. ViTs are utilized in various applications, including image recognition, image segmentation, and so on.

Given an input image I with dimensions $H \times W \times C$, where H is the height, W is the width, and C is the number of channels (e.g., 3 for RGB): The image is divided into N patches, each of size $P \times P$, resulting in patches with dimensions $\frac{H}{P} \times \frac{W}{P} \times C$. Each patch is then flattened and linearly projected to a D-dimensional embedding vector using a learnable matrix $E \in \mathbb{R}^{(P^2 \cdot C)\times D}$. Positional embedding is then added to the patch embeddings to retain positional information. The final input embeddings for the transformer encoder is the sum of patch embedding and position embedding. The input embeddings are passed through a series of transformer encoder layers, each consisting of multi-head self-attention (MHSA) and feed-forward (FF) networks. The output of the transformer encoder is then used for various vision tasks, such as classification, segmentation, etc.

CLIP CLIP [367] is designed to understand and generate associations between textual and visual data. The model is trained on a large dataset of images and their corresponding textual descriptions, allowing it to learn a wide range of visual concepts and their linguistic representations.

CLIP consists of two encoders: an image encoder and a text encoder. The goal of CLIP is to learn a joint embedding space where corresponding images and texts are close to each other. Let I be an input image. The image encoder f_I maps I to a D-dimensional embedding vector:

$$v = f_I(I) \in \mathbb{R}^D \tag{2.10}$$

Let T be an input text (e.g., a caption or a sentence). The text encoder f_T maps T to a D-dimensional embedding vector:

$$w = f_T(T) \in \mathbb{R}^D \tag{2.11}$$

CLIP uses a contrastive loss to train the encoders. For a batch of N image-text pairs, the contrastive loss is defined as:

$$L = -\sum_{i=1}^{N} \log \frac{\exp(v_i \cdot w_i / \tau)}{\sum_{j=1}^{N} \exp(v_i \cdot w_j / \tau)} \tag{2.12}$$

where τ is a temperature parameter, and $v_i \cdot w_j$ is the dot product between the image embedding v_i and the text embedding w_j. The goal of the training is to minimize the contrastive loss, which encourages the model to align the embeddings of corresponding image-text pairs while pushing apart the embeddings of non-corresponding pairs.

Part I

Learning Cross-modal Semantic Alignment

Chapter 3

Multimodal Summarization via Cross-domain Alignment

In this chapter, we start with the discussion of learning cross-domain alignment, with a focus on a new multimedia application named Multimodal summarization with multimodal output (MSMO). MSMO is a recently explored application in language grounding. It plays an essential role in real-world applications, i.e., automatically generating cover images and titles for news articles or providing introductions to online videos. However, existing methods extract features from the whole video and article and use fusion methods to select the representative one, thus usually ignoring the critical structure and varying semantics with video/document. In this chapter, we propose a Semantics-Consistent Cross-domain Summarization (SCCS) model based on optimal transport alignment with visual and textual segmentation. Our method first decomposes both videos and articles into segments in order to capture the structural semantics, and then follows a cross-domain alignment objective with optimal transport distance, which leverages multimodal interaction to match and select the visual and textual summary. We evaluate our method on three MSMO datasets, and achieved performance improvement by 8% & 6% of textual and 6.6% &5.7% of video summarization, respectively, which demonstrated the effectiveness of our method in producing high-quality multimodal summaries.

3.1 Introduction

New multimedia content in the form of short videos and corresponding text articles has become a significant trend in influential digital media. This popular media type has been shown to be successful in drawing users' attention and delivering essential information in an efficient manner. MSMO has recently drawn increasing attention. Different from traditional video or textual summarization [143, 191], where the generated summary is either a keyframe or textual description, MSMO aims at producing both visual and textual summaries simultaneously, making this task more complicated. Previous works addressed the MSMO task by processing the whole video and the whole article together which overlooked the structure and semantics of different domains [96, 114, 115, 150, 295, 399, 576].

The video and article can be regarded as being composed of several topics related to the main idea, while each topic specifically corresponds to one sub-idea. Thus, treating the whole

Figure 3.1: Comparison with previous work: We proposed a segment-level cross-domain alignment model to preserve the structural semantics consistency within two domains for the MSMO task. We solve an optimal transport problem to optimize the cross-domain distance, which in turn finds the optimal match.

video or article uniformly and learning a general representation ignores these structural semantics and easily leads to biased summarization. To address this problem, instead of learning averaged representations for the whole video & article, we focus on exploiting the original underlying structure. The comparison of our approach and previous works is illustrated in Figure 3.1. Our model first decomposes the video & article into segments to discover the content structure, then explores the cross-domain semantics relationship at the segment level. We believe this is a promising approach to exploit the consistency lie in the structural semantics between different domains.

Previous models applied attention or fusion mechanisms to compute image-text relevance scores, finding the best match of the sentences/images within the whole document/video, regardless of the context, which used one domain as an anchor. However, an outstanding anchor has more weight in selecting the corresponding pair. To overcome this, we believe the semantics structure is a crucial characteristic that can not be ignored. Based on this hypothesis, we propose Semantics-Consistent Cross-domain Summarization (SCCS), which explores segment-level cross-domain representations through OT-based multimodal alignment to generate both visual and textual summaries. We decompose the video/document into segments based on its semantic structure, then generate sub-summaries of each segment as candidates. We select the final summary from these candidates instead of a global search, so all candidates are in a fair competition arena.

Our contributions can be summarized as follows :

- We propose SCCS, a segment-level alignment model for MSMO tasks.
- Our method preserves the structural semantics and explores the cross-domain relationship through optimal transport to match and select the visual and textual summary.
- On three datasets, our method outperforms baselines in both textual and video summarization results qualitatively and quantitatively.

Figure 3.2: SCCS at work: A real example of the summarization process given by our SCCS method. Here we conduct OT-based cross-domain alignment to each keyframe-sentence pair, and a smaller OT distance means better alignment. (For example, the best-aligned text and image summary (0.08) delivers the flooding content clearly and comprehensively.)

- Our method serves as a hierarchical MSMO framework and provides better interpretability via OT alignment. The OT coupling shows sparse patterns and specific temporal structure for the embedding vectors of ground-truth-matched video and text segments, providing interpretable learned representations.

Since MSMO generates both visual & textual summaries, We believe the optimal summary comes from the video and text pair that are both 1) semantically consistent, and 2) best matched globally in a cross-domain fashion. In addition, our framework is more computationally efficient as it conducts cross-domain alignment at the segment level instead of inputting whole videos/articles.

3.2 Related Work

Multimodal Summarization Multimodal summarization explored multiple modalities, i.e., audio signals, video captions, transcripts, video titles, etc, for summary generation. [114, 324, 494, 541] learned the relevance or mapping in the latent space between different modalities. In addition to only generating visual summaries, [19, 236, 576] generated textual summaries by taking audio, transcripts, or documents as input along with videos or images, using seq2seq model [445] or attention mechanism [24]. Recent trending on the MSMO task has also drawn much attention [114, 115, 295, 552, 576].

Optimal Transport OT studies the geometry of probability spaces [471], a formalism for finding and quantifying mass movement from one probability distribution to another. OT defines the Wasserstein metric between probability distributions, revealing a canonical geometric structure with rich properties to be exploited. The earliest contribution to OT originated from Monge in the eighteenth century. Kantorovich rediscovered it under a different formalism, namely the Linear Programming formulation of OT. With the development of scalable solvers, OT is widely applied to many real-world problems and applications [11, 54, 55, 96, 105, 215, 230, 540, 573].

3.3 Proposed Method

SCCS is a segment-level cross-domain semantics alignment model for the MSMO task, where MSMO aims at generating both visual and language summaries. We follow the problem setting in