Bridging Multimedia Modalities: Enhanced Multimodal AI Understanding and Intelligent Agents

Sushant Gautam
sushant@simula.com
Department of Holistic Systems (HOST)
Simula Metropolitan Center for Digital Engineering (SimulaMet)
Oslo, Norway

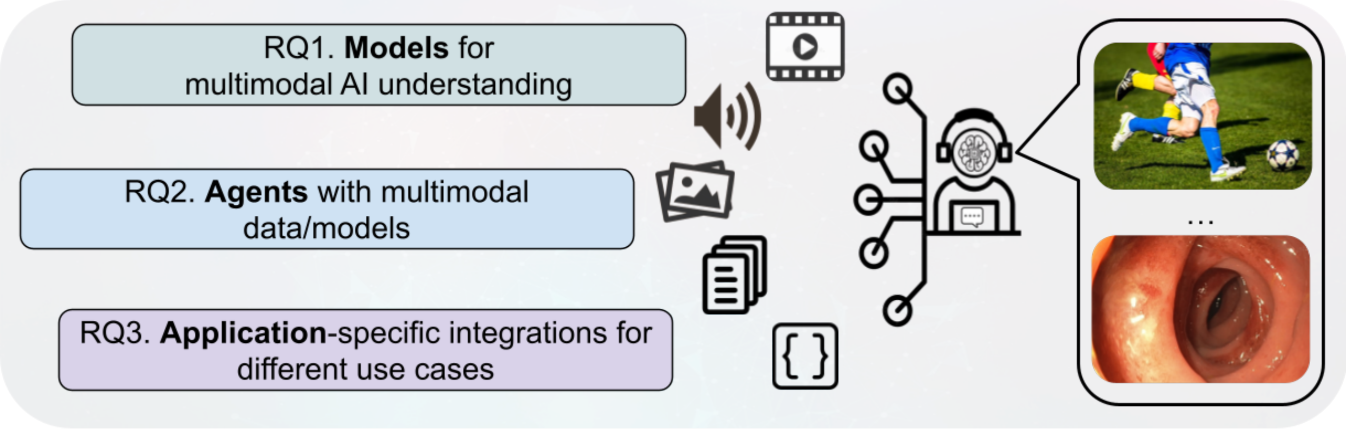
RQ1. **Models** for multimodal AI understanding
RQ2. **Agents** with multimodal data/models
RQ3. **Application**-specific integrations for different use cases
...

Figure 1: Overview of the proposed research through key research questions (RQ).

ABSTRACT

With the increasing availability of multimodal data, especially in the sports and medical domains, there is growing interest in developing Artificial Intelligence (AI) models capable of comprehending the world in a more holistic manner. Nevertheless, various challenges exist in multimodal understanding, including the integration of multiple modalities and the resolution of semantic gaps between them. The proposed research aims to leverage multiple input modalities for the multimodal understanding of AI models, enhancing their reasoning, generation, and intelligent behavior. The research objectives focus on developing novel methods for multimodal AI, integrating them into conversational agents with optimizations for domain-specific requirements. The research methodology encompasses literature review, data curation, model development and implementation, evaluation and performance analysis, domain-specific applications, and documentation and reporting. Ethical considerations will be thoroughly addressed, and a comprehensive research plan is outlined to provide guidance. The research contributes to the field of multimodal AI understanding and the advancement of sophisticated AI systems by experimenting with multimodal data to enhance the performance of state-of-the-art neural networks.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Natural language processing**; **Knowledge representation and reasoning**; • **Information systems** → **Multimedia and multimodal retrieval**; • **Human-centered computing** → **User centered design**.

KEYWORDS

Multimodal Fusion; Multimedia; AI Understanding; Conversational Agents

ACM Reference Format:
Sushant Gautam. 2023. Bridging Multimedia Modalities: Enhanced Multimodal AI Understanding and Intelligent Agents. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 9–13, 2023, Paris, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3577190.3614225

1 INTRODUCTION AND MOTIVATION

Multimodal understanding refers to the capability of AI systems to simultaneously process information from multiple modalities (text, speech, images, etc.), leading to a more comprehensive understanding of the content. For instance, in soccer game summarization, multimodal understanding involves analyzing both visual cues (such as player movements and goals) and accompanying audio (including crowd cheers and commentator's remarks) to generate an

ICMI '23, October 9–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10.
https://doi.org/10.1145/3577190.3614225

informative summary. Similarly, in medical image understanding, combining information from different imaging modalities (such as X-rays and MRI scans) with textual medical reports could yield a more accurate diagnosis and a comprehensive assessment of a patient's condition.

Recent years have witnessed a growing interest in AI for multimodal understanding [17, 20, 29, 38]. This interest stems from the increased availability of multimodal data [35], including sports videos (comprising image frames and audio) paired with text commentaries [21] and medical images/videos coupled with clinical characteristics [30]. Leveraging multimodal data, AI models can achieve a more holistic comprehension of the world [1, 17, 23]. Developing robust AI models for multimodal understanding presents several challenges. One such challenge is addressing the semantic gap between diverse modalities [4, 6, 29, 38]. For instance, a captivating soccer moment, like a player kicking the ball into the net, can be described textually as "a shot by a player toward the net." Yet, there is no direct mapping between the visual features of game elements and situations on the field and the descriptive words employed.

Another significant challenge is effectively integrating multiple modalities to capture their complementary information [2]. This entails meticulous design of model architecture and training processes, along with a profound understanding of inter-modality relationships [37]. Despite these challenges, substantial progress has been made in AI for multimodal understanding in recent years [17]. Novel techniques have emerged to bridge the semantic gap between distinct modalities [17], contributing to the training of AI models that demonstrate impressive performance across a range of multimodal tasks. This research aims to forge connections between diverse modalities in multimodal understanding by developing innovative models to enhance such understanding and seamlessly integrating them into domain-specific optimized autonomous agents.

The subsequent sections of this document are organized as follows: In Section 2, we provide a concise background and overview of related work. In Section 3, we delve into our research questions. Section 4 outlines our methodology to address these research questions and discusses our dissemination plan. Preliminary results from experiments pertaining to the research questions are presented in Section 5. Finally, Section 6 concludes the document with notes on the next steps.

2 BACKGROUND AND RELATED WORK

The integration of multimodal approaches and language models has driven significant advancements in AI systems. Gao et al. [8] demonstrated that their multimodal feature-based method achieved higher accuracy and effectiveness in recognizing human motion during soccer games. Similarly, Wu et al. [34] proposed a multimodal two-stream 3D network framework that improved recognition performance through the synergistic use of complementary multimodal information. Extensive experiments on challenging action recognition datasets substantiated their approach. These studies [8, 22, 34] collectively underscore the potential of multimodal approaches in enhancing AI understanding within multimedia analysis.

The Meta-Transformer [39] stands as a multimodal framework enabling unified learning across 12 diverse modalities without paired training data. This approach leverages a frozen encoder to extract high-level semantic features from raw input data, exhibiting efficacy across various tasks and applications. However, limitations include high computational complexity, absence of temporal and structural awareness, and potential constraints for cross-modal generation tasks. MultiBench [16] offers a comprehensive and standardized large-scale benchmark for multimodal learning. Encompassing diverse datasets, modalities, and prediction tasks, MultiBench accelerates progress and enhances real-world robustness in multimodal AI research.

Previous research on AI-based soccer game summarization pipelines [9–11] focused on automating the generation of comprehensive textual summaries for soccer games. These pipelines incorporated multimodal inputs such as video and audio streams, alongside accessible game metadata, to produce variable-length game summaries. Notable achievements included fine-tuning a Longformer model [3] for generating game summaries based on textual game captions. Additionally, these studies explored the use of game audio to prioritize events for summary inclusion, utilizing the Root Mean Square (RMS) audio intensity score to assess event importance [10]. To address the challenge of comprehending information-rich videos, researchers are increasingly exploring direct input of temporal image frames [22, 25, 31, 33], aiming to enhance multimodal AI understanding across diverse modalities [38].

Researchers from various domains are actively exploring the realm of deep learning for multimodal data fusion. Gao et al. [7] delve into pioneering models that fuse diverse data types, offering fundamental insights specifically in the context of multimodal big data. Stahlschmidt et al. [28] prioritize nonlinear biomedical data fusion, proposing a taxonomy to enhance fusion strategy selection within the biomedical domain. To complement this landscape, Lipkova et al.[36] enriches the discourse by providing a nuanced exploration of AI's integration into the multifaceted world of oncology data. Moreover, the work of Li et al. [14] systematically reviews deep learning's role in remote sensing data fusion, shedding light on emerging trends in the domain of multimodal remote sensing data fusion.

Moreover, interest has surged in agent-based frameworks to enhance language model capabilities, providing modular components and pre-built chains for various tasks [12, 13, 26, 27]. These chains consist of modular components that are customizable and seamlessly integrable with multimodal data sources and functions. Notable examples include frameworks developed by Kraus et al. [13], Shen et al. [26], and Shridhar et al. [27].

LAnguage Model Analysis (LAMA) [24] extensively explores the inherent relational knowledge in pretrained language models, highlighting their potential as unsupervised open-domain question answering systems. Liu et al. [19] provide a comprehensive overview of prompt-based learning in natural language processing, discussing advantages, mathematical notations, and various dimensions such as pre-trained language models, prompts, and tuning strategies. Ding et al. [5] introduce Open-Prompt, a comprehensive toolkit for prompt-based learning with Pre-trained Language Models (PLMs)s, offering efficiency, modularity, and adaptability to various Natural Language Processing (NLP) tasks. Liu et al. [18] present a rigorous comparison between few-shot In-Context Learning (ICL) and Parameter-efficient Fine-tuning (PEFT) methods, showcasing

PEFT's efficient prompt-based approach, superior accuracy, and reduced computational costs. Video-ChatGPT [25] combines video-adapted visual encoding with Large Language Models (LLMs), enabling human-like conversations about videos. Video-LLaMA [38] empowers LLMs to comprehend both visual and auditory content in videos, addressing challenges in capturing temporal changes in visual scenes and integrating audio-visual signals, making it a promising prototype for audio-visual conversational agents.

The proposed research aims to build upon these achievements by bridging the gap between different input modalities, including those addressed through computer vision, audio analysis, and language models. This will enhance AI understanding for downstream tasks, including but not limited to generation, classification, and detection. Such enhancements will also improve conversational agents, making them more intelligent by leveraging multimodal data sources and functions. By addressing open challenges and exploring novel directions, this research has the potential to significantly advance multimodal understanding and contribute to the development of more sophisticated AI systems.

3 RESEARCH QUESTIONS

Building upon the accomplishments of previous research, the proposed study aims to synthesize various modalities in multimodal understanding. This is achieved through the development of novel models for enhanced multimodal understanding, their integration into conversational agents with domain-specific optimizations, and the addressing of application requirements and performance concerns. The research is guided by the following research question:

Main Research Question (RQ):

How can existing AI models be adapted or extended to handle multimodal data and enhance reasoning and generation capabilities through the effective representation and alignment of different input modalities? How can these adapted models be optimized for integration into conversational agents, addressing domain-specific application requirements for various use cases?

Sub-Research Questions:

Research Question 1 (RQ1): *How can existing AI models be adapted or extended to handle multimodal data and enhance reasoning and generation capabilities through the effective representation and alignment of different input modalities?* In RQ1, the focus is on modifying and combining pre-existing AI models, originally designed for individual modalities, to efficiently process and align information from multiple sources. By integrating knowledge and advancements from unimodal domains, the goal is to develop novel multimodal AI models capable of reasoning and generating content by effectively fusing information from diverse input sources. The successful development of such adapted models will contribute to enhancing multimodal AI understanding.

Research Question 2 (RQ2): *Can the integration of multimodal data and models in conversational agents enhance their capabilities and reflect improved intelligent behavior?* RQ2 aims to investigate how leveraging multimodal information can empower conversational agents to gain contextual understanding, generate richer responses, and make better decisions. By fusing data from various modalities, conversational agents can become more contextually aware and provide more informative and personalized interactions, ultimately improving their overall performance and user experience.

Research Question 3 (RQ3): *How can the incorporation of multimodal data and models be optimized to address domain-specific application requirements and performance concerns for different use cases, such as sports and healthcare?* RQ3 addresses the optimization of incorporating multimodal data and models to meet domain-specific application requirements and performance concerns in different use cases, such as sports and healthcare. The motivation behind this question lies in the need to tailor multimodal understanding approaches to specific domains, ensuring their effectiveness and efficiency in practical applications.

The fusion of different modalities, as explored in RQ1, holds the potential to enhance the quality and richness of generation across various domains [15, 38]. For example, it can improve game summaries by capturing the nuances of gameplay, player actions, and visual cues. In the medical domain, challenges such as privacy concerns and limited data availability for machine learning training often necessitate the use of synthetic data. Conversely, the sports domain benefits from easily accessible long game videos, publicly available statistics, and news data, facilitating multimodal understanding tasks within this domain. These variations in modalities and data availability contribute to the distinct considerations and approaches required for each use case.

4 METHODOLOGY AND APPROACH

The proposed research constitutes a vital component of a 3-year (36-month) PhD program. The outlined plans below detail how the PhD student aims to acquire knowledge, practical skills, and contribute significantly to the field of multimodal understanding.

4.1 Research Plan

Literature Review (RQ1, RQ2, RQ3): The research commences with an extensive review of existing literature on multimodal understanding, the alignment of diverse data modalities, fusion techniques, integration of pre-trained models, and the amalgamation of multimodal data and models within autonomous agents. This analysis identifies strengths, limitations, and research gaps, forming the foundation for the development of novel multimodal models that represent and align various input modalities effectively. The literature review also examines studies addressing domain-specific application requirements and performance concerns, offering insights into optimizing multimodal data and model integration for various use cases.

Data Curation (RQ1, RQ2, RQ3): Acknowledging limitations in existing datasets, particularly the lack of unified and comprehensive data sources for multimodal understanding [9, 11, 17], the research will explore, curate, and enhance datasets. Ensuring diversity and representation of multimodal data while aligning with domain-specific requirements in sports and healthcare, these curated datasets will undergo rigorous reprocessing to ensure quality, consistency, and augmented training and evaluation efficiency.

Model Development and Implementation (RQ1): To address RQ1, the research will explore novel methods for multimodal understanding and fusion. These methods aim to effectively represent and align diverse input modalities, including visual, audio, and textual data. Integration of pre-trained models, such as Convolutional Neural Networks (CNNs), Transformers, and language models, will be investigated to enhance multimodal understanding and improve generation quality. Developed methods and algorithms will be implemented using suitable programming languages and frameworks, considering the specific requirements of diverse data modalities and models.

Integration with Conversational Agents (RQ2): Aligned with RQ2, the research will explore methods to enhance the capabilities of autonomous agents and reflect improved intelligent behavior by leveraging multimodal data and models. This integration involves adapting multimodal models to interface with conversational agents, enabling the agents to process and interpret various input modalities. The integrated system's performance will be evaluated through experiments and objective/subjective studies, assessing the enhanced capabilities and improved intelligent behaviors exhibited by the agents.

Domain-specific Applications (RQ3): The efficacy of multimodal understanding within developed models and methods will be evaluated through practical domain-specific applications, such as sports and healthcare, addressing RQ3. Models will be fine-tuned based on domain-specific datasets and application scenarios, addressing challenges unique to each domain. Techniques to enhance system efficiency, reduce computational complexity, and improve real-time performance will be explored to optimize multimodal model integration.

Evaluation and Performance Analysis (RQ1, RQ2, RQ3): Rigorous evaluation of proposed models and methods will measure performance in terms of reasoning accuracy, generation quality, and intelligent behavior. Evaluations will involve comparing results with baseline models and existing approaches, as well as subjective assessment through human feedback. The proposed methods and system will be validated using benchmark datasets and real-world scenarios to ensure generalizability, effectiveness, and applicability beyond the research problem.

Documentation and Reporting (RQ1, RQ2, RQ3): The entire research process, methodologies, experimental setups, findings, and insights will be documented and disseminated through research papers, reports, presentations, and open-source code implementations. Sharing curated datasets and code promotes reproducibility and facilitates further research efforts.

5 PRELIMINARY RESULTS

As part of this research, we have begun experimenting with various multimodal datasets from different domains and exploring emerging model architectures. In our initial experiment, we utilized Video-LLaMA [38], a platform enhanced by large language models with video and audio understanding capabilities [32], and performed fine-tuning using the SoccerNet-Captions dataset [21]. The concept is illustrated in Figure 2.

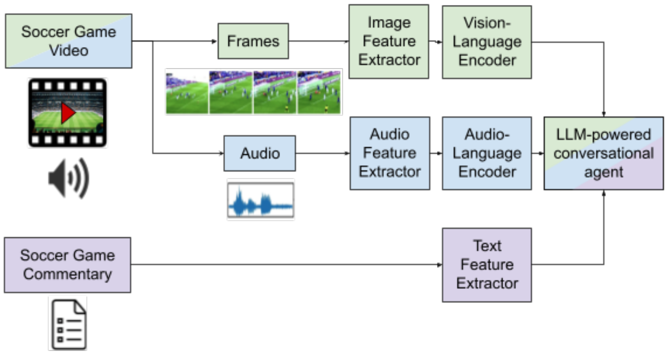

Figure 2: LLM-powered conversational agent leveraging video frames (images), audio, and game commentary (text) for enhanced understanding.

To start, soccer game videos from the SoccerNet dataset are divided into uniform-length (T seconds) video chunks, each covering a game event with corresponding audio commentary from the SoccerNet-Captions dataset (beginning at time point E). These chunks are strategically generated to span a time duration starting b seconds before and ending a seconds after the commentary for the game event, such that $(E + a) - (E - b) = T$.

Segmenting the videos in this manner equips the fine-tuning process with pairs of video chunks and commentary. This approach trains the model to comprehend the contextual context surrounding specific soccer game events. Video-LLaMA thus learns the intricate relationship between visual cues within the video, the associated audio content, and the textual commentary that describes the event. The video chunks provide visual information that enables the model to analyze the game's dynamics and player movements, while the corresponding audio complements this by capturing ambient sounds, crowd reactions, and other auditory cues that contribute to a comprehensive understanding of the event.

Currently, the fine-tuning of the vision-language branch has been completed, and efforts are ongoing in fine-tuning the audio-language branch. We are actively working on evaluating the outputs, and this evaluation process is a work in progress.

6 CONCLUSION AND NEXT STEPS

This proposed research aims to leverage multiple input modalities to enhance reasoning, generation, and intelligent behavior in conversational agents, thereby advancing multimodal understanding and integration. The research objectives encompass the development of novel **models** for enhanced multimodal understanding through the fusion of diverse modalities, their integration into conversational **agents**, and the pursuit of domain-specific optimizations to address **application** requirements and performance concerns in sports and healthcare domains. The proposed methodology includes comprehensive research, progress, and dissemination plans, ensuring accessibility and high impact. Preliminary results from ongoing work illustrate the potential of experimenting with various multimodal datasets and emerging model architectures, specifically focusing on enhancing automated video and audio understanding of soccer game events in the sports domain. The forthcoming steps will be guided by the research questions outlined in Section 3 and the research and dissemination plans detailed in Section 4.1.

REFERENCES

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 2 (Jan. 2018), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

[2] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. 2022. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *Vis. Comput.* 38, 8 (Aug. 2022), 2939–2970. https://doi.org/10.1007/s00371-021-02166-7

[3] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv* (April 2020). https://doi.org/10.48550/arXiv.2004.05150 arXiv:2004.05150

[4] Yanbei Chen, Yongqin Xian, A Koepke, Ying Shan, and Zeynep Akata. 2021. Distilling audio-visual knowledge by compositional contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 7016–7025.

[5] Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. 2022. OpenPrompt: An Open-source Framework for Prompt-learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* Association for Computational Linguistics, Dublin, Ireland, 105–113. https://doi.org/10.18653/v1/2022.acl-demo.10

[6] Junhao Feng, Guohua Wang, Changmeng Zheng, Yi Cai, Ze Fu, Yaowei Wang, Xiao-Yong Wei, and Qing Li. 2023. Towards Bridged Vision and Language: Learning Cross-modal Knowledge Representation for Relation Extraction. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[7] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. 2020. A Survey on Deep Learning for Multimodal Data Fusion. *Neural Comput.* 32, 5 (May 2020), 829–864. https://doi.org/10.1162/neco_a_01273

[8] Yuzhou Gao. 2021. Human Motion Recognition Based on Multimodal Characteristics of Learning Quality in Football Scene. *Mathematical Problems in Engineering* (2021). https://doi.org/10.1155/2021/7963616

[9] Sushant Gautam. 2022. AI-based Soccer Game Summarization: From Video Highlights to Dynamic Text Summaries. Master's Thesis, Tribhuvan University. https://www.researchgate.net/publication/363857936_AI-based_Soccer_Game_Summarization_From_Video_Highlights_to_Dynamic_Text_Summaries

[10] Sushant Gautam, Cise Midoglu, Saeed Shafiee Sabet, Dinesh Baniya Kshatri, and Pål Halvorsen. 2022. Assisting Soccer Game Summarization via Audio Intensity Analysis of Game Highlights. In *Proceedings of 12th IOE Graduate Conference*, Vol. 12. Institute of Engineering, Tribhuvan University, Nepal, 25 – 32. https://conference.ioe.edu.np/publications/ioegc12/IOEGC-12-004-12009.pdf

[11] Sushant Gautam, Cise Midoglu, Saeed Shafiee Sabet, Dinesh Baniya Kshatri, and Pål Halvorsen. 2022. Soccer Game Summarization using Audio Commentary, Metadata, and Captions. In *NarSUM '22: Proceedings of the 1st Workshop on User-centric Narrative Summarization of Long Videos.* Association for Computing Machinery, New York, NY, USA, 13–22. https://doi.org/10.1145/3552463.3557019

[12] Significant Gravitas. 2023. Auto-GPT. https://github.com/Significant-Gravitas/Auto-GPT.

[13] Mathias Kraus, Julia Anna Bingler, Markus Leippold, Tobias Schimanski, Chiara Colesanti Senni, Dominik Stammbach, Saeid Ashraf Vaghefi, and Nicolas Webersinke. 2023. Enhancing Large Language Models with Climate Resources. *arXiv* (March 2023). https://doi.org/10.48550/arXiv.2304.00116 arXiv:2304.00116

[14] Jiaxin Li, Danfeng Hong, Lianru Gao, Jing Yao, Ke Zheng, Bing Zhang, and Jocelyn Chanussot. 2022. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *Int. J. Appl. Earth Obs. Geoinf.* 112 (Aug. 2022), 102926. https://doi.org/10.1016/j.jag.2022.102926

[15] Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. 2021. Bridging Text and Video: A Universal Multimodal Transformer for Audio-Visual Scene-Aware Dialog. *IEEE/ACM Trans. Audio Speech Lang. Process.* 29 (2021), 2476–2483. https://doi.org/10.1109/TASLP.2021.3065823

[16] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. 2021. MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1).*

[17] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2022. Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. *arXiv* (Sept. 2022). https://doi.org/10.48550/arXiv.2209.03430 arXiv:2209.03430

[18] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 1950–1965. https://proceedings.neurips.cc/paper_files/paper/2022/file/0cde695b83bd186c1fd456302888454c-Paper-Conference.pdf

[19] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* 55, 9 (Jan. 2023), 1–35. https://doi.org/10.1145/3560815

[20] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation. *arXiv* (Feb. 2020). https://doi.org/10.48550/arXiv.2002.06353 arXiv:2002.06353

[21] Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. 2023. SoccerNet-Caption: Dense Video Captioning for Soccer Broadcasts Commentaries. *arXiv* abs/2304.04565 (2023). https://doi.org/10.48550/arXiv.2304.04565 arXiv:2304.04565

[22] Salman Khan Muhammad Maaz, Hanoona Rasheed and Fahad Khan. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. *ArXiv 2306.05424* (2023).

[23] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *ICML*.

[24] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases?. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Association for Computational Linguistics, Hong Kong, China, 2463–2473. https://doi.org/10.18653/v1/D19-1250

[25] Madeline C. Schiappa, Yogesh S. Rawat, and Mubarak Shah. 2023. Self-Supervised Learning for Videos: A Survey. *ACM Comput. Surv.* 55, 13s (July 2023), 1–37. https://doi.org/10.1145/3577925

[26] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. *arXiv* (March 2023). https://doi.org/10.48550/arXiv.2303.17580 arXiv:2303.17580

[27] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. 2023. Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation. In *Proceedings of The 6th Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 205)*, Karen Liu, Dana Kulic, and Jeff Ichnowski (Eds.). PMLR, 785–799. https://proceedings.mlr.press/v205/shridhar23a.html

[28] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. 2022. Multimodal deep learning for biomedical data fusion: a review. *Briefings Bioinf.* 23, 2 (March 2022), bbab569. https://doi.org/10.1093/bib/bbab569

[29] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2023. Any-to-Any Generation via Composable Diffusion. *arXiv* (May 2023). https://doi.org/10.48550/arXiv.2305.11846 arXiv:2305.11846

[30] Vajira Thambawita, Steven A. Hicks, Andrea M. Storås, Thu Nguyen, Jorunn M. Andersen, Oliwia Witczak, Trine B. Haugen, Hugo L. Hammer, Pål Halvorsen, and Michael A. Riegler. 2023. VISEM-Tracking, a human spermatozoa tracking dataset. *Sci. Data* 10, 260 (May 2023), 1–8. https://doi.org/10.1038/s41597-023-02173-4

[31] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems* 35 (2022), 10078–10093.

[32] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv* (Feb. 2023). https://doi.org/10.48550/arXiv.2302.13971 arXiv:2302.13971

[33] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 14549–14560.

[34] Hanbo Wu. 2021. Spatiotemporal Multimodal Learning With 3D CNNs for Video Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 32 (2021), 1250–1261. https://doi.org/10.1109/tcsvt.2021.3077512

[35] Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. 2023. Cap4Video: What Can Auxiliary Captions Do for Text-Video Retrieval?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 10704–10713.

[36] Peng Xu, Xiatian Zhu, and David A Clifton. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[37] Duoyi Zhang, Richi Nayak, and Md Abul Bashar. 2021. Exploring Fusion Strategies in Deep Learning Models for Multi-Modal Classification. In *Data Mining.* Springer, Singapore, 102–117. https://doi.org/10.1007/978-981-16-8531-6_8

[38] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. *arXiv* (June 2023). https://doi.org/10.48550/arXiv.2306.02858 arXiv:2306.02858

[39] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. 2023. Meta-Transformer: A Unified Framework for Multimodal Learning. *arXiv preprint arXiv:2307.10802* (2023).