A Survey of Multimodal Learning and Its Applications

www.surveyx.cn

Abstract

Multimodal learning represents a significant advancement in artificial intelligence, integrating diverse data modalities—text, images, audio, and video—to overcome the limitations of unimodal systems. This survey explores key methodologies such as the Multimodal Learning with Alternating (MLA) approach and Dual-Modality Prompt Tuning (DPT), which enhance model adaptability and performance across various datasets. In speech recognition, incorporating global context has improved ASR systems, while frameworks like BIKE and II-CLVM demonstrate state-of-the-art performance in video recognition and video-music retrieval tasks, respectively. The potential of multimodal learning in audio representation is emphasized, with models like HighMMT achieving superior results in high-modality scenarios. In medical imaging, adapting CLIP models enhances image-text alignment, improving diagnostic capabilities. The MXM-CLR framework and COTS model significantly advance cross-modal representation learning, achieving state-of-the-art results in retrieval tasks. The survey highlights the transformative potential of multimodal learning in enhancing AI systems' performance and robustness. Future research should focus on refining training strategies, integrating emerging techniques such as generative models, and addressing data scarcity and quality issues. By exploring these avenues, multimodal learning can continue to drive significant advancements in AI, enhancing its applicability across diverse real-world scenarios.

1 Introduction

1.1 Significance of Multimodal Learning

Multimodal learning (MML) represents a significant advancement in artificial intelligence by integrating diverse data types—text, images, audio, and video—to enhance understanding and prediction across various domains [1]. This approach effectively addresses the heterogeneity gap between modalities, a challenge that traditional methods, which often convert data into a common vector space, struggle to overcome [2]. By leveraging the complementary strengths of different sensory modalities, MML enhances the comprehensiveness and accuracy of machine learning models, overcoming the limitations of unimodal systems [3].

The incorporation of external knowledge sources into vision-and-language (VL) representation learning highlights MML's transformative potential in bridging knowledge gaps in VL tasks, particularly in scenarios with limited labeled multimodal data [4]. Understanding modality interactions can lead to improved reasoning and task performance.

In automated speech recognition (ASR), the fusion of audio and visual data significantly enhances performance, especially in noisy environments [5]. This integration is critical for developing robust systems capable of functioning effectively under challenging conditions. Additionally, video information can substantially improve audio representation learning, which is essential for various audio-related tasks [6].

The theoretical and computational foundations of MML are continually evolving, with ongoing research identifying common themes and open questions in the field [7]. This evolution drives the

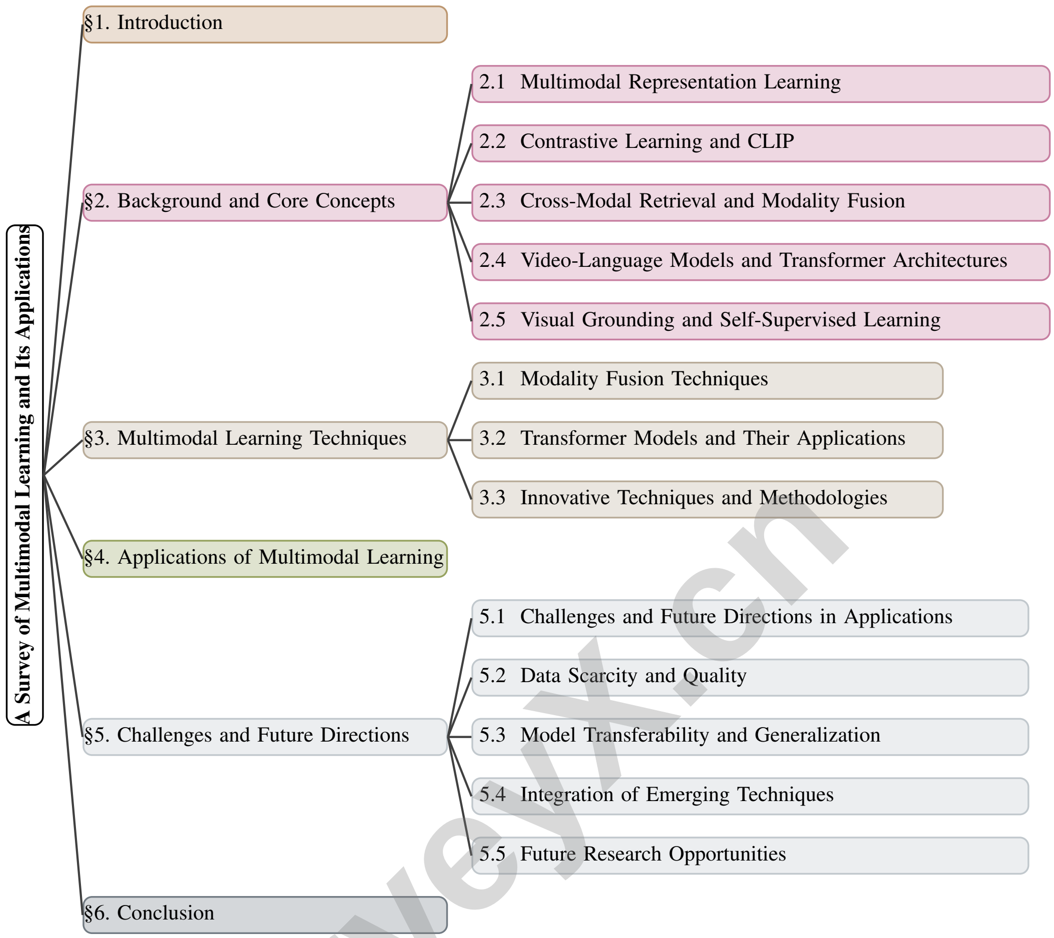


Figure 1: chapter structure

development of sophisticated multimodal systems that address the complexities of real-world applications. The integration of multiple modalities not only enhances the performance and interpretability of machine learning systems but also broadens their applicability across diverse domains, making it a cornerstone of modern artificial intelligence research [8].

1.2 Motivation for the Survey

This survey aims to bridge significant knowledge gaps in multimodal deep learning (MMDL) by providing a comprehensive review of diverse modalities and their applications [1]. Despite rapid advancements in unimodal systems, integrating multiple data types remains a formidable challenge, necessitating innovative approaches and methodologies [3]. The survey explores these advancements, focusing on modality heterogeneity, connections, and interactions, which are essential for understanding multimodal machine learning's foundational elements [7].

A key motivation for this survey is to address modality bias in existing video question-answering benchmarks, which limits the assessment of models' abilities to integrate information from diverse sources [9]. By examining these biases, the survey seeks to enhance evaluation frameworks and improve the robustness of multimodal models.

Furthermore, the survey fills literature gaps by exploring the interplay of various modalities and their technical designs, contributing to a deeper understanding of multimodal machine learning [8]. This exploration is particularly relevant for vision-and-language (VL) models, where comprehension of commonsense, factual, and temporal knowledge is critical for real-world application performance [4].

This survey provides an exhaustive overview of multimodal learning, addressing challenges and advancements in co-learning and offering insights into effective integration strategies. By systematically analyzing the integration of modalities—such as text and images—and addressing challenges like knowledge gaps and alignment issues, it aims to establish a framework for future research and applications in diverse domains including social media analysis, medical imaging, and emotion recognition [10, 11, 4].

1.3 Relevance in Current Research Landscape

The multimodal learning landscape has transformed significantly, with recent research emphasizing its critical role in enhancing model performance across applications like emotion recognition and video captioning [12]. This progress is propelled by the development of Multimodal Large Language Models (MLLMs), which have substantially improved the processing and integration of multimodal data, underscoring their relevance amid ongoing research and technological advancements [13].

Understanding the theoretical foundations of multimodal learning is crucial for navigating the complexities associated with diverse data integration. Recent studies highlight that a robust grasp of these foundations is essential for advancing research in this domain [14]. MML plays a pivotal role in bridging semantic and heterogeneity gaps, particularly between image and text data, vital for seamless integration across modalities [10].

In educational contexts, analyzing multimodal data—such as speech, video, and eye gaze—has deepened our understanding of learner behaviors and outcomes. This integration is vital for developing educational technologies that respond more effectively to learners' nuanced needs [15]. Additionally, significant strides have been made in modeling multimodal interactions, leading to improved performance in critical applications such as healthcare and affective computing [7].

The significance of multimodal learning is further underscored by its potential to advance the field towards achieving human-like intelligence in machines. By enabling machines to process and interpret information similarly to humans, multimodal learning paves the way for sophisticated AI systems capable of executing complex tasks with greater accuracy and efficiency [8]. As research evolves, the integration of multimodal learning techniques remains a cornerstone in the pursuit of advanced and capable artificial intelligence systems.

1.4 Structure of the Survey

This survey is meticulously structured to provide a comprehensive exploration of multimodal learning, beginning with an introduction that underscores the significance and motivation behind this field of study. The introduction sets the stage by discussing the relevance of multimodal learning in the current research landscape, highlighting its transformative impact across various domains. Following this, the paper delves into foundational concepts in the "Background and Core Concepts" section, where key theories and methodologies are examined, including CLIP, Whisper, cross-modal retrieval, modality fusion, video-language models, contrastive learning, transformer models, visual grounding, and self-supervised learning.

The third section, "Multimodal Learning Techniques," presents an in-depth analysis of various methodologies employed in the field, with a particular focus on self-supervised learning. This section explores modality fusion techniques and the application of transformer models, alongside innovative approaches propelling the field forward. Subsequently, the survey transitions into "Applications of Multimodal Learning," where the practical implementations of these techniques are explored across domains such as healthcare, video and audio processing, sentiment analysis, and real-world scenarios.

In the penultimate section, "Challenges and Future Directions," the survey addresses current obstacles in multimodal learning, including data scarcity, model transferability, and the integration of emerging techniques. This section also identifies potential future research opportunities that could further advance the field. Finally, the survey concludes with a summary of key findings, reflecting on the overarching importance of multimodal learning and its potential to revolutionize various fields. This structured approach ensures a holistic understanding of multimodal learning, providing valuable insights for researchers and practitioners alike.The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Multimodal Representation Learning

Multimodal representation learning synthesizes different sensory modalities—text, images, audio, and video—into a unified framework to improve task comprehension and performance, leveraging their complementary strengths and overcoming unimodal limitations [3]. A key challenge is learning complex intra-modal and cross-modal interactions while maintaining robustness against missing or noisy modalities [16]. Techniques such as AV-BERT, which integrates audio and visual data via masked language modeling, enhance automatic speech recognition in challenging environments [5]. Dual-modality prompt tuning (DPT) further refines focus on target visual concepts by adjusting text and visual prompts [17].

Recent surveys categorize research based on challenges and methods in multimodal co-learning, providing a comprehensive overview [3]. This taxonomy aids in understanding effective integration strategies. Additionally, exploring adversarial robustness in Vision-Language Pre-training (VLP) models highlights the need for a unified framework to generate adversarial text-image pairs for robust multimodal systems [18]. Contrastive learning methods like CURVES enhance multimodal interaction by producing graphical utterances denoting visual referents, bridging the semantic gap between modalities [19]. Benchmarks for learning discriminative audio representations from multiple formats address challenges in integrating raw audio with spectral representations [20].

2.2 Contrastive Learning and CLIP

Contrastive learning is crucial in multimodal learning, enhancing integration and differentiation of data modalities by leveraging inherent similarities and differences, thereby addressing modality competition challenges [21]. This technique aligns representations across modalities, such as images and text, enhancing multimodal system robustness [21]. The CLIP model exemplifies contrastive learning, achieving notable performance in image-text retrieval and zero-shot classification using extensive datasets [22]. However, challenges like partial false negatives in cross-modal contrastive learning can hinder model optimization [23], and improper minimization of mutual information involving negative samples may lead to semantic information loss [19].

Innovative methods address these challenges. For instance, CLIPC generates composite image-caption pairs to enhance vision-language model robustness [22]. The Multi-View Contrastive learning framework models intra-modal and inter-modal correlations, addressing limitations of single-view contrastive learning in VLP models [21]. Theoretical foundations of contrastive learning in multimodal settings emphasize accurately estimating mutual information in high-dimensional spaces, elucidating modality relationships [23]. By focusing on positive and negative pairs, training dynamics of multimodal contrastive learning models are enhanced, improving alignment and balance [19].

2.3 Cross-Modal Retrieval and Modality Fusion

Cross-modal retrieval and modality fusion are critical in bridging semantic and distributional gaps between diverse data modalities like text, images, audio, and video. Cross-modal retrieval involves retrieving semantically relevant information from one modality using a query from another, addressing challenges posed by inconsistent distributions and representations [24]. The heterogeneous nature of data complicates effective comparison and retrieval, particularly between audio and visual modalities [24]. The primary obstacle is the heterogeneity gap, complicating the mapping relationship between modalities and leading to ineffective cross-modal learning [23].

Models like COTS enhance image-text retrieval by integrating multiple levels of cross-modal interactions [22]. However, existing cross-modal contrastive representation learning methods often underutilize multifold observations, resulting in suboptimal data instance representation [19]. Employing semantically composite examples during pretraining improves performance in cross-modal retrieval tasks, enhancing models like CLIP [22].

Modality fusion aims to create unified representations by combining diverse data types, addressing inefficiencies from independent processing [24]. The challenge of aligning and integrating modalities during training and testing necessitates advanced strategies for robust model performance [23]. The complexity of calculating similarity between different modalities and understanding its influence on decision-making processes underscores the intricacy of modality fusion [19].

Despite these challenges, modality fusion is essential for robust multimodal systems, facilitating the alignment and integration of diverse modalities to enhance AI performance. The ongoing development of advanced techniques in cross-modal retrieval and modality fusion is pivotal in advancing multimodal learning, enabling AI systems to process and interpret complex multimodal information more accurately and efficiently [22]. By addressing representation, alignment, and quantification challenges, these techniques significantly contribute to the evolution of multimodal learning in real-world applications.

2.4 Video-Language Models and Transformer Architectures

Video-language models and transformer architectures are crucial in advancing multimodal learning, enhancing understanding and generation of video content through language. These models leverage both video and textual data, enabling robust interpretations across modalities. The AudioVisual Recurrent Network (AVRN) exemplifies the integration of audio and visual features for enhanced video summarization, underscoring joint feature utilization's importance in video-language tasks [25].

Current methodologies include explicit and implicit alignment methods and various fusion frameworks like kernel-based and attention-based approaches [11]. These frameworks facilitate seamless integration of video and text modalities, crucial for tasks such as Video Question Answering (VideoQA). Benchmarks for VideoQA tasks evaluate models' joint understanding and representation capabilities [26].

Transformer-based architectures have proven effective in multimodal contexts, as recent surveys document their evolution and impact [27]. The High Modality Multimodal Transformer (HighMMT) employs cross-modal attention mechanisms to learn robust multimodal representations, enhancing integration of information from various modalities [28]. This approach highlights transformers' transformative potential in processing and integrating diverse data types.

Innovative methods like Bidirectional Cross-Modal Knowledge Exploration (BIKE) leverage Video-to-Text and Text-to-Video knowledge transfer, enhancing video recognition capabilities [29]. Additionally, the Domain-Agnostic Multi-Modal Video Retrieval (DAMMVR) method integrates audio and video features through a novel transformer architecture, demonstrating transformers' versatility in complex multimodal retrieval tasks [30].

2.5 Visual Grounding and Self-Supervised Learning

Visual grounding and self-supervised learning are pivotal in advancing multimodal frameworks by enhancing alignment and integration of diverse data modalities. Visual grounding associates linguistic descriptions with specific elements in visual content, crucial for developing context-aware AI systems. This process addresses challenges like spurious correlations within AI models, refining visual-language tasks' accuracy [31]. Integrating a multi-layered architecture with differential learning rates enhances image-to-text transformation processes, illustrating visual grounding's potential in multimodal applications [32]. Advanced entropy estimation techniques, such as those in InfoMeter, provide insights into 3D object detection performance, highlighting mutual information's role in improving visual grounding [33].

Self-supervised learning leverages data's intrinsic structure to generate supervisory signals, enabling models to learn useful representations without extensive labeled datasets. This approach is particularly beneficial in multimodal settings where labeled data are scarce or costly. Self-supervised learning techniques have been applied successfully in audio-visual contexts, enhancing long-form content understanding like movies through audio-visual integration [34]. In automatic speech recognition, masked language modeling in self-supervised frameworks creates global, multi-modal encodings, improving performance [5].

The synergy between visual grounding and self-supervised learning is exemplified by contrastive learning methods, which differentiate inputs to understand relationships between visual and textual elements [35]. A volume-based alignment measure, replacing traditional cosine similarity, allows for simultaneous geometric alignment of all modalities, enriching semantic information and enhancing multimodal representation learning [36]. Additionally, methods like CLIPArTT utilize automatic text prompt construction during inference to create pseudo-labels, facilitating test-time adaptation and enhancing multimodal systems' robustness [37].

3 Multimodal Learning Techniques

The integration of diverse data modalities—text, images, audio, and video—is pivotal in multimodal learning for developing robust AI systems. This integration enhances model accuracy by utilizing complementary information and facilitates knowledge transfer in data-scarce scenarios. As illustrated in Figure 2, this figure presents a hierarchical classification of multimodal learning techniques, emphasizing modality fusion, transformer applications, and innovative methodologies. It showcases the significant advancements in AI systems' capabilities to process and interpret complex multimodal information. Recent advancements in multimodal alignment and fusion techniques have significantly improved machine learning models' capabilities in complex tasks across domains such as social media analysis, medical imaging, and emotion recognition. However, challenges like alignment issues, noise resilience, and feature representation disparities persist, necessitating ongoing research to optimize multimodal systems for scalability, robustness, and generalizability [15, 11, 3, 38]. This exploration begins with modality fusion techniques, foundational in synthesizing diverse data forms, paving the way for complex interactions and applications in multimodal learning.

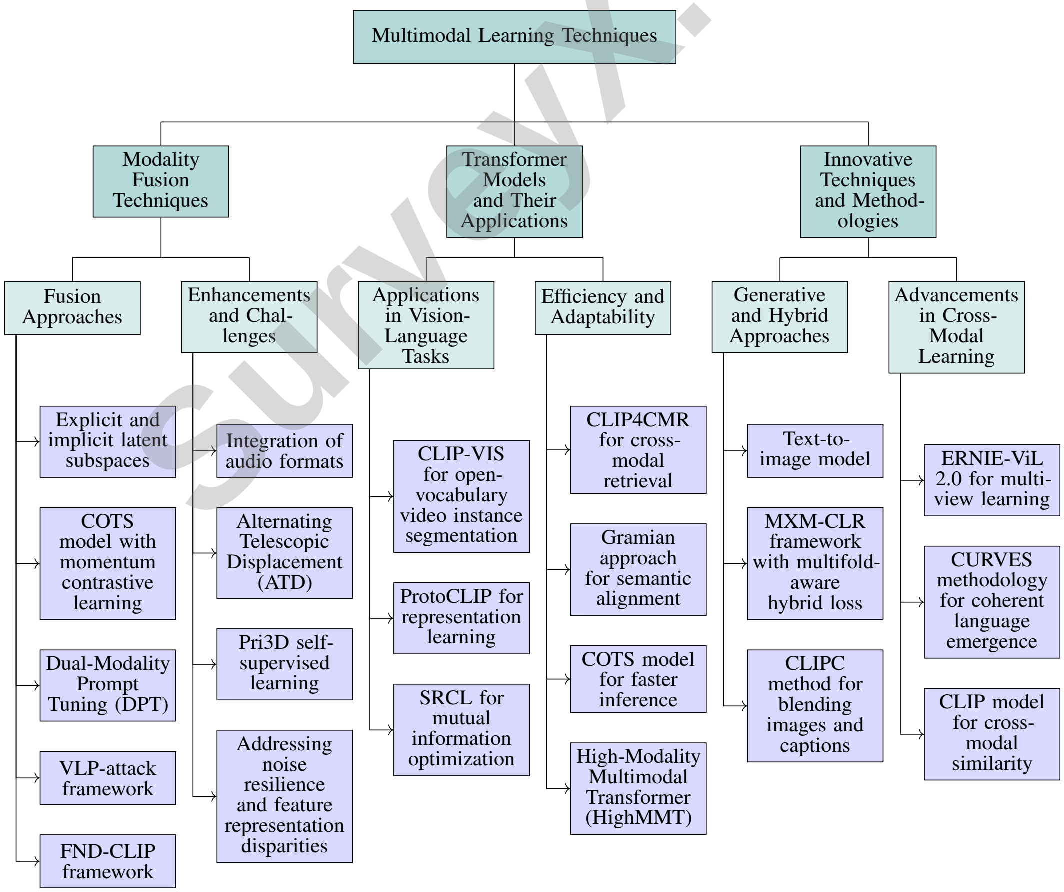


Figure 2: This figure presents a hierarchical classification of multimodal learning techniques, emphasizing modality fusion, transformer applications, and innovative methodologies. It illustrates the integration of diverse data modalities, showcasing significant advancements in AI systems' capabilities to process and interpret complex multimodal information.

AI-generated, for reference only.

3.1 Modality Fusion Techniques

Modality fusion techniques are essential for effectively integrating diverse data types, such as audio, text, and video, to enhance model performance and robustness. These techniques create unified representations that leverage the complementary strengths of different modalities, capturing intricate interrelations. A significant approach involves using explicit and implicit latent subspaces to preserve critical semantic information from both modalities, thus enhancing the retrieval process [39].

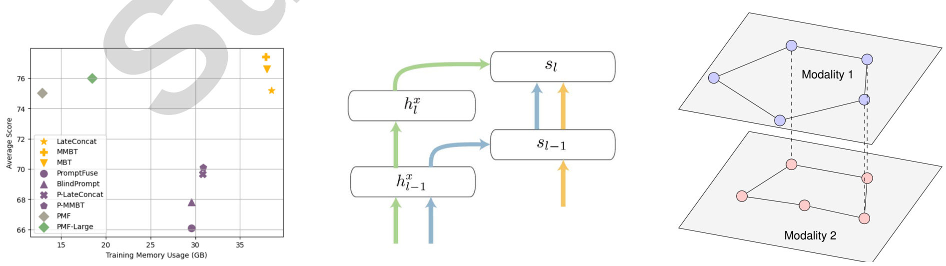
The COTS (Collaborative Two-Stream) model exemplifies advanced modality fusion by employing instance-level interaction through momentum contrastive learning and enhancing token-level interaction with masked vision-language modeling [40]. Additionally, Dual-Modality Prompt Tuning (DPT) dynamically generates class-aware visual prompts alongside text prompts, improving the fusion of visual and textual data [17].

In adversarial contexts, the VLP-attack framework integrates adversarial text and image generation into a unified modality fusion technique, showcasing the potential of multimodal information integration to enhance robustness [18]. Furthermore, the FND-CLIP framework utilizes the pretrained CLIP model to extract and align features from text and images, employing a modality-wise attention mechanism to enhance classification accuracy in multimodal fake news detection [41].

The integration of various audio formats, akin to image processing augmentations, has been shown to enhance learning, emphasizing the importance of diverse modality representations [20]. The Alternating Telescopic Displacement (ATD) approach systematically shifts and expands feature representations across modalities to address alignment challenges [2].

Pri3D, a self-supervised contrastive learning approach utilizing RGB images and registered point clouds, illustrates the potential of modality fusion by learning visual representations without human annotations [23]. This underscores modality fusion's transformative potential in enhancing AI capabilities across applications.

Modality fusion techniques significantly advance multimodal learning by enabling AI systems to integrate and interpret diverse data types—text, images, audio, and video—thereby improving accuracy and efficiency. These techniques align complementary information across modalities, enhancing model performance in various applications, including social media analysis, medical imaging, and emotion recognition. They also address challenges such as noise resilience and feature representation disparities, leading to more robust and generalizable AI systems capable of processing complex multimodal information [15, 11, 42]. The advancements in modality fusion highlight its transformative potential in enhancing AI capabilities across diverse applications.



(a) Comparison of Average Scores and Training Memory Usage for Different Models[43]

(b) A neural network architecture with multiple layers and connections[42]

(c) Two modalities of data visualization[11]

Figure 3: Examples of Modality Fusion Techniques

As illustrated in Figure 3, the integration of diverse data modalities is pivotal for enhancing model performance and efficiency. The first subfigure presents a scatter plot comparing the average scores and training memory usage of different models, revealing trade-offs between performance and resource consumption. Models such as "LateConcat" and "MMBT" are distinguished by unique symbols and colors, highlighting diversity in architecture and efficiency. The second subfigure delves into a neural network architecture, emphasizing the layered structure and connectivity essential for effective modality fusion. The third subfigure visualizes two modalities of data on distinct planes, interconnected to depict relationships, underscoring the importance of combining multiple data

AI-generated, for reference only.

sources to enrich the learning process and develop robust models. This figure encapsulates modality fusion techniques and the multifaceted strategies employed in multimodal learning [43, 42, 11].

3.2 Transformer Models and Their Applications

Transformer models have become instrumental in advancing multimodal learning due to their robust capacity to handle sequential data and capture complex dependencies across modalities. In vision-language tasks, these models have been adapted for various applications, including video instance segmentation and cross-modal retrieval. The CLIP-VIS model exemplifies this adaptation by utilizing the frozen CLIP model for open-vocabulary video instance segmentation, enabling instance segmentation and tracking without category annotations [44]. This flexibility highlights transformer architectures' capability in processing diverse data types.

In multimodal pretraining, ProtoCLIP employs prototypes to guide representation learning, enhancing the efficiency of representation grouping [45]. The SRCL method optimizes mutual information between image/text anchors and their negative counterparts, applying a similarity-based weight to the negatives, thereby improving the alignment and balance of multimodal representations [46].

The CLIP4CMR framework utilizes the CLIP model as a backbone for supervised cross-modal retrieval, generating common representations for image and text data [47]. This approach bridges the semantic gap between modalities, enhancing retrieval tasks. Additionally, the Gramian approach captures semantic alignment by minimizing the volume of a parallelotope formed by modality embeddings, ensuring closely aligned modalities yield a smaller volume [36].

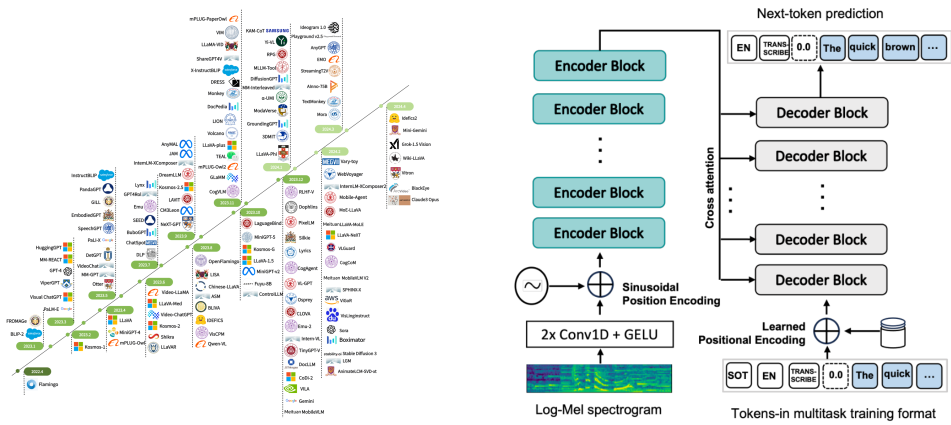
The efficiency of transformer models in multimodal contexts is further demonstrated by the COTS model, which achieves significantly faster inference speeds compared to single-stream models while outperforming two-stream models in retrieval tasks [40]. This efficiency is crucial for real-time applications requiring rapid processing and integration of multimodal information. The use of a loss function that considers multiple pairwise combinations of input modalities facilitates effective cross-modal representation learning, showcasing transformers' adaptability in diverse multimodal contexts [48].

Transformers are increasingly essential in advancing multimodal learning, enabling seamless integration and interpretation of diverse data types, including text, images, audio, and sensor inputs. Recent developments, such as the High-Modality Multimodal Transformer (HighMMT) and the Meta-Transformer framework, demonstrate significant improvements in representation learning by quantifying modality and interaction heterogeneity. These innovations allow for efficient processing and feature extraction across diverse modalities without extensive paired training data. Consequently, transformer architectures are proving highly effective in enhancing performance across various multimodal tasks, including visual question answering, natural language processing, and data mining applications [49, 27, 28]. Through innovative adaptations and advanced learning techniques, transformers remain at the forefront of multimodal research, driving progress across numerous applications.

As shown in Figure 4, multimodal learning techniques and transformer models are pivotal innovations driving significant advancements across various applications. The first part of the example, "Timeline of Major Language Model Projects," visually represents the evolution of language model projects from April 2022 to April 2024, underscoring the rapid development and deployment of language models, each marked by distinct logos and release years. The second part, "Next-Task Prediction with Sinusoidal Position Encoding," delves into neural networks designed for next-token prediction, emphasizing sinusoidal position encoding within the architecture, which features multiple encoder and decoder blocks enhanced by cross-attention mechanisms. This network efficiently generates sequences of tokens, demonstrating transformer models' transformative capabilities in handling complex prediction tasks. Together, these examples encapsulate the profound influence of transformer models in modern machine learning, highlighting their multifaceted applications and ongoing evolution [13, 50].

3.3 Innovative Techniques and Methodologies

Recent advancements in multimodal learning have introduced various innovative techniques and methodologies that enhance the integration and interaction of diverse data modalities. A notable



(a) Timeline of Major Language Model Projects[13]

(b) Next-Task Prediction with Sinusoidal Position Encoding[50]

Figure 4: Examples of Transformer Models and Their Applications

innovation is the use of generative transformer models to create synthetic visual data, exemplified by the text-to-image model, which contrasts with prior methods focused on data removal or imputation without generative capabilities [51]. This highlights the potential of generative models to enrich existing multimodal frameworks by generating new, meaningful data representations.

The MXM-CLR framework introduces a multifold-aware hybrid loss (MFH), allowing the method to consider multiple positive observations when calculating relationships, thereby overcoming limitations of previous single-fold-oriented methods [52]. This innovation underscores the importance of considering multiple perspectives to enhance the robustness and accuracy of multimodal representations.

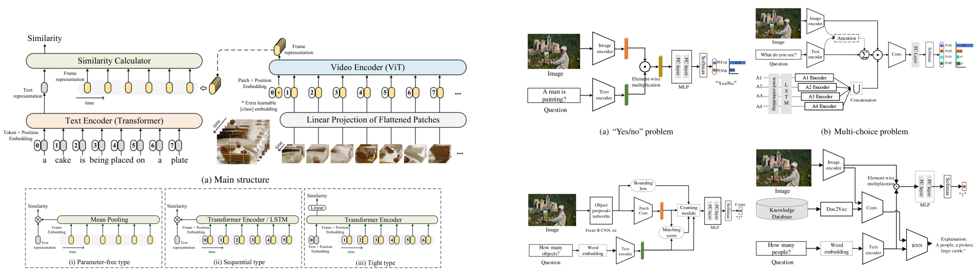
In cross-modal learning, the CLIPC method exemplifies innovation by creating new training examples through blending images and merging captions, improving model performance on downstream tasks [22]. The ERNIE-ViL 2.0 model introduces a multi-view learning framework that constructs various visual/textual views, significantly improving the robustness and generalization of cross-modal representations [21].

The CURVES methodology advances the field by facilitating the emergence of a coherent language that generalizes across different contexts and referent compositions, outperforming traditional methods reliant on discrete tokens [19]. This highlights the potential of contrastive learning in generating versatile and adaptable multimodal representations.

In medical imaging, future research directions suggest developing more efficient training methods, exploring cross-modal applications, and enhancing the generalizability of CLIP models in diverse medical contexts [53]. These directions emphasize the need for continuous innovation to address specific challenges posed by complex multimodal data in specialized domains.

The use of the CLIP model to measure cross-modal similarity and guide the feature fusion process represents a significant advancement over existing methods that treat modalities separately, as demonstrated in multimodal fake news detection [41]. This underscores the importance of integrated modality treatment for improving the accuracy and reliability of multimodal systems.

Innovative methodologies are advancing multimodal learning by enhancing AI systems' capabilities to accurately and efficiently process and interpret complex multimodal information. Recent technological developments have improved the collection and analysis of diverse data types, such as speech, video, and eye gaze, leading to a more comprehensive understanding of learning and training environments. This includes the introduction of novel frameworks and taxonomies categorizing multimodal approaches, as well as advanced techniques like mid fusion and citation graph pruning, refining data analysis. Emerging multimodal deep learning models, such as auto-encoders and generative adversarial networks, facilitate both uni-directional and bidirectional tasks, further enriching content understanding. Despite these advancements, ongoing research is essential to bridge the gap between multimodal applications and foundational AI principles, ensuring that the subtleties revealed by multimodal data are effectively integrated into AI systems [10, 15]. By addressing core challenges and leveraging novel techniques, these advancements propel the field forward, offering new possibilities for developing sophisticated AI systems.



(a) Video-Text Similarity Calculation with Transformer Encoder[54]

(b) Deep Learning Models for Image-Text Question Answering[10]

Figure 5: Examples of Innovative Techniques and Methodologies

As shown in Figure 5, innovative techniques and methodologies are pivotal in advancing AI systems' capabilities to process and understand diverse data types in multimodal learning. The first example focuses on video-text similarity calculation using a Transformer encoder, which employs separate text and video encoders to transform sequences of words and frames into hidden states, facilitating similarity determination between textual and visual content. The second example illustrates deep learning models for image-text question answering, showcasing distinct models for binary "Yes/no" questions and multi-choice questions. These models demonstrate deep learning's potential to interpret and respond to complex queries by analyzing both visual and textual inputs, pushing the boundaries of machine understanding and interaction in multimodal contexts. Together, these examples underscore the significance of innovative methodologies in developing sophisticated AI systems capable of processing and integrating information from multiple sources [54, 10].

4 Applications of Multimodal Learning

4.1 Healthcare and Medical Imaging

In healthcare, multimodal learning integrates varied data types to enhance diagnostics and patient care. Techniques like MedCLIP leverage contrastive learning to align medical images with textual data, improving efficiency and interpretability in medical contexts [55]. Models such as eCLIP excel in cross-modal retrieval, outperforming traditional methods by retrieving relevant medical images through textual queries [56]. Multimodal Large Language Models (MLLMs) extend capabilities to complex medical data, including Electronic Health Records (EHR), supporting comprehensive healthcare solutions [13]. PubMedCLIP enhances Visual Question Answering (VQA) systems by focusing on pathological features, improving medical image interpretation [57, 58]. Frameworks like CoMM demonstrate robustness across healthcare datasets, underscoring their applicability in real-world scenarios [59].

4.2 Video and Audio Processing

Multimodal learning revolutionizes video and audio processing by integrating diverse modalities for enhanced retrieval and classification. The Wav2CLIP model exemplifies this by aligning audio with text for improved classification and retrieval [60]. Techniques like GRACE and CoAVT highlight superior performance in audio-visual interactions [46, 11]. ModalityMirror and LLM2CLIP frameworks enhance audio classification and multimodal representation [61, 62]. CMMixer captures interdependencies between audio and video, improving retrieval performance [63]. Fast adaptation of contrastive models facilitates efficient learning in video question answering and text-to-video retrieval [6]. Experiments on datasets like Clotho V2, AudioCaps, MSR-VTT, and VATEX illustrate the enhanced capabilities of multimodal approaches in video and audio processing [6, 62]. CLIP-powered models further demonstrate versatility across complex tasks [46].

4.3 Sentiment Analysis and Emotion Recognition

Multimodal learning advances sentiment analysis and emotion recognition by leveraging multiple data types for enhanced accuracy. Models evaluated on emotion detection datasets demonstrate the efficacy of integrating text, audio, and visual data [64]. Frameworks like Transformative Robust Multimodal Learning (TRML) address challenges of missing data, ensuring robust performance [65]. Innovative models such as iCODE show adaptability in sentiment analysis [66]. Multimodal routing techniques improve interpretability by capturing relationships between emotional cues [67]. In spoken language tasks, models like Whisper enhance understanding through audio-text integration [50]. These advancements underscore the potential of multimodal learning to improve sentiment analysis across languages and contexts [10, 64, 38].

4.4 Multimodal Learning in Real-World Scenarios

Multimodal learning demonstrates significant potential in real-world applications by integrating diverse data types to enhance performance. The AVDCNN model improves speech enhancement through audio-visual integration, beneficial in challenging environments [68]. FLEX-CLIP excels in cross-modal retrieval, outperforming state-of-the-art methods in various scenarios [69]. Harmonized multimodal Gaussian Process Latent Variable Models (GPLVM) and trimodal models advance social media content classification and retrieval [70, 71]. Multimodal learning's ability to adapt to distribution shifts indicates its broader applicability in vision-language tasks [72]. Techniques like MS-CLIP and APLe enhance efficiency and generalization in zero-shot recognition and domain-shift scenarios [73, 74]. The GTI-MM model demonstrates robustness in data-scarce environments, outperforming existing methods [51]. By integrating text, images, audio, and video, multimodal learning enables advanced tasks like cross-modal retrieval and emotion recognition, improving performance in applications from crisis detection to education. Ongoing developments promise to address challenges and drive future innovations in AI [64, 10, 1, 27, 15].

5 Challenges and Future Directions

The domain of multimodal learning encompasses numerous challenges and future directions, pivotal for the effective application of these techniques. Key issues include the integration of diverse data sources and ensuring model robustness. This section delves into specific challenges and underscores critical future research areas to enhance the efficacy and adaptability of multimodal systems.

5.1 Challenges and Future Directions in Applications

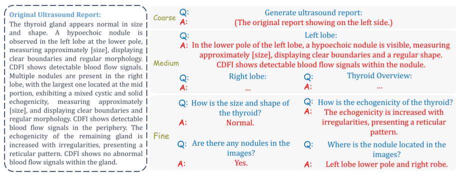
Multimodal learning encounters significant hurdles in co-learning scenarios, especially with missing modalities or high noise levels [3]. Current research often falls short in addressing these complexities, pointing to the need for robust models to overcome these adversities. The discrete nature of text data complicates the generation of transferable multimodal adversarial samples, essential for enhancing adversarial robustness [18].

Future research should explore advanced fusion techniques, such as attention mechanisms, to capture local features within implicit spaces more effectively [39]. Hybrid knowledge integration approaches could improve model generalizability across vision-language tasks, addressing dataset quality and model adaptability limitations [4].

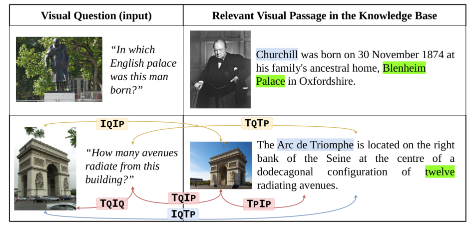
Developing explainable fake news detection systems is another pertinent direction, focusing on elucidating classification reasoning and identifying suspicious elements in news articles [41]. Moreover, incorporating intra-modality relationships and extending frameworks like MXM-CLR to other modalities, such as video data, can enhance the robustness and applicability of multimodal systems [52].

In multimodal adversarial learning, generating transferable adversarial samples remains challenging, particularly due to the discrete nature of text data, necessitating innovative approaches that bridge modality gaps [18]. The ATD method shows promise by improving accuracy in multimodal tasks while reducing computational complexity [2].

Future research should also explore cross-modal alignment implications in self-supervised learning contexts, where multiple sensing modalities are fused [23]. Addressing these challenges and pursuing these research directions can propel multimodal learning towards more effective, scalable, and adaptable solutions.



(a) The image shows a table with three columns and five rows, each containing a question and an answer.[58]



(b) Visual Question Answering with Knowledge Base[75]

Figure 6: Examples of Challenges and Future Directions in Applications

As illustrated in Figure 6, examples of "Challenges and Future Directions in Applications" provide insights into the complexities and advancements in visual question answering (VQA) and data interpretation. The first example features a structured table titled "Original Ultrasound Report," highlighting the challenge of conveying detailed medical information. The second example showcases a VQA scenario using a knowledge base, illustrating the integration of visual data with contextual knowledge to enhance accuracy in VQA systems. These examples underscore ongoing challenges in developing systems that can seamlessly integrate and interpret complex information.

5.2 Data Scarcity and Quality

Data scarcity and quality significantly challenge multimodal learning, impacting the development of robust models. A primary concern is the reliance on large labeled datasets, often scarce in specialized domains like medical imaging, where high-quality annotations are essential [53]. This scarcity is compounded by the need for multi-scale features and specialized knowledge.

The quality of available data is another critical issue, as models frequently overfit to noise, diminishing their generalization capabilities [63]. In contrastive learning frameworks, performance heavily relies on the quality of constructed views; poor-quality views can lead to suboptimal outcomes [21]. Benchmarks often fail to capture the complexities of audio-text retrieval due to data scarcity and challenges in aligning diverse captions with audio clips, underscoring the need for comprehensive evaluation frameworks [24].

In scenarios with highly dissimilar modalities, existing methods struggle to model complex relationships, especially when faced with missing or noisy inputs. The AV-BERT method exemplifies the challenge of data scarcity, as it relies on visual data availability during training [5]. Furthermore, the limited effectiveness of vision-language models in low-data contexts highlights the need for approaches that leverage data diversity [22].

Efforts to address these challenges include updating outdated benchmarks and carefully selecting clusters to ensure effective training [34]. By improving data availability and quality, and developing methodologies that minimize reliance on large annotated datasets, the field can progress towards more robust and adaptable solutions.

5.3 Model Transferability and Generalization

Model transferability and generalization are pivotal challenges in multimodal learning, where adapting models across diverse tasks and datasets is crucial for robust performance. A primary obstacle is the discrepancy between input and output spaces across different modalities, complicating the transfer of learned knowledge and necessitating effective alignment of representation spaces [76]. This challenge is exacerbated by dependency on the quality and diversity of training data, as models may struggle to generalize effectively with insufficiently diverse datasets [32].

The Normalized Contrastive Learning (NCL) framework exemplifies difficulties in generalizing across different query distributions, where deviations between training and testing distributions can

AI-generated, for reference only.

impact performance [77]. Similarly, the MoRE framework highlights challenges in applying learned representations beyond initially tested datasets, indicating limitations in adaptability [78].

Continual learning algorithms face the challenge of maintaining zero-shot capabilities while fine-tuning on new tasks, essential for preserving versatility and avoiding catastrophic forgetting [79]. The MoVA framework addresses these challenges by enhancing generalization through adaptive expert selection, reducing bias from irrelevant vision encoders and improving performance across diverse tasks [80].

The Dual-Modality Prompt Tuning (DPT) approach highlights difficulties in modifying visual features extracted by image encoders, limiting current methods' applicability to downstream tasks [17]. Additionally, the COTS model faces computational challenges associated with calculating similarity scores for all query-candidate pairs, underscoring the trade-off between computational efficiency and model interaction [40].

Parameter sharing in multimodal models, while beneficial for performance, may not fully capture unique modality characteristics, crucial for effective transferability [73]. Furthermore, balancing modalities during training is critical, as imbalances can lead to suboptimal expert utilization and performance drops [81].

5.4 Integration of Emerging Techniques

Integrating emerging techniques in multimodal learning is essential for addressing current challenges and enhancing AI system robustness. One promising direction involves refining training strategies to better accommodate multimodal data complexities. For instance, the Sugar framework has shown potential in improving model performance across various contexts [82]. Investigating such frameworks can lead to more effective approaches for integrating and leveraging diverse data types.

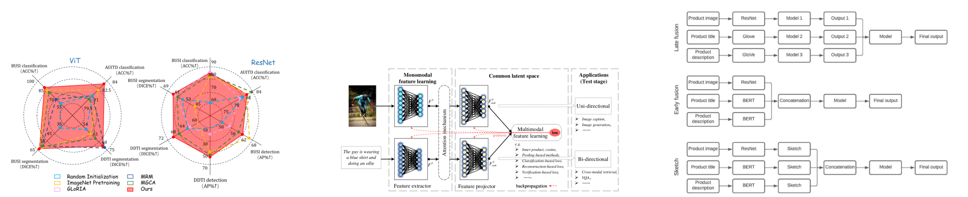
Incorporating novel methods like generative models, which create synthetic visual data, offers a transformative approach to enriching multimodal frameworks, contrasting with traditional methods that focus on data removal or imputation [51]. Additionally, developing hybrid loss functions that consider multiple positive observations, as seen in MXM-CLR, enhances multimodal representation robustness by overcoming limitations of single-fold-oriented methods [52].

The application of attention mechanisms and hybrid knowledge integration approaches further improves model generalizability across diverse vision-language tasks, addressing current dataset quality and model adaptability limitations [4]. Moreover, advanced entropy estimation techniques, such as those in InfoMeter, provide insights into complex task performance, highlighting mutual information's role in improving visual grounding and overall accuracy [33].

Emerging methodologies, such as the Alternating Telescopic Displacement (ATD) method, demonstrate improved accuracy in multimodal tasks by optimizing feature alignment and reducing computational complexity [2]. These advancements underscore the potential of optimized methodologies in advancing multimodal learning, offering new possibilities for sophisticated AI system development.

The integration of emerging techniques is crucial for addressing prevalent challenges such as data alignment, noise resilience, and representation disparities while enhancing accuracy and applicability across diverse applications. By leveraging complementary information from multiple modalities—text, images, audio, and video—researchers can improve content understanding and facilitate knowledge transfer, particularly in scenarios with limited data. This multifaceted approach drives progress in the field and reveals intricate patterns that single-modality analyses may overlook, paving the way for future advancements in multimodal co-learning and feature embedding strategies [10, 15, 11, 3]. By leveraging novel approaches and refining existing methodologies, researchers can enhance multimodal systems' capabilities, enabling them to process and interpret complex data with greater accuracy and efficiency.

As shown in Figure 7, the integration of emerging techniques in machine learning and computer vision represents an evolving frontier characterized by various challenges and future directions. The first example compares Vision Transformer (ViT) and ResNet models in image classification and segmentation tasks, highlighting evolving performance evaluation metrics. The second example emphasizes the significance of integrating monomodal and multimodal feature extraction processes. Lastly, the exploration of three distinct fusion strategies for product image and description analysis reveals the complexity and versatility of processing multimodal inputs. Collectively, these examples demonstrate



(a) Comparison of ViT and ResNet Models in Image Classification and Segmentation Tasks[58]

(b) Multimodal Feature Learning and Its Applications[10]

(c) Three Different Fusion Strategies for Product Image and Description Analysis[83]

Figure 7: Examples of Integration of Emerging Techniques

current capabilities of emerging techniques and point towards future research opportunities aimed at refining and expanding their applicability.

5.5 Future Research Opportunities

Future research in multimodal learning offers numerous opportunities to enhance AI systems' robustness, efficiency, and adaptability by addressing current limitations and exploring innovative methodologies. A significant focus area is developing unified models capable of effectively learning from multiple modalities while optimizing pretraining strategies to enhance performance and reduce computational costs [8]. This involves refining gradient modification mechanisms and test-time dynamic fusion to bolster model robustness in complex multimodal scenarios [3].

Exploring local dimension implications on model performance and developing robust multimodal architectures that effectively bridge modality gaps are crucial for advancing the field [84]. The application of theoretical insights from contrastive loss functions to more complex models and investigating additional training techniques to enhance multimodal learning are promising research directions [85]. Additionally, creating more complementary questions that require genuine integration of multiple modalities can advance understanding of modality interactions [9].

Research should also refine parameter grouping strategies and apply models like HighMMT to increasingly diverse modalities and tasks, enhancing adaptability and performance across varied contexts [28]. Furthermore, enhancing model robustness against dataset biases and exploring additional modalities for improved retrieval are vital areas for further investigation [30].

Future work could focus on improving accuracy in interaction quantification and exploring self-supervised learning applications, leading to more effective multimodal frameworks [86]. Exploring unsupervised or semi-supervised learning techniques can enhance multimodal approaches' applicability in scenarios with limited labeled data [16].

Refinements in negative sampling strategies and applying proposed methods to other media forms beyond movies offer additional avenues for advancing multimodal learning [34]. Moreover, future research should further refine the GRAM method, exploring its application to more diverse and complex multimodal datasets while enhancing robustness against misalignment [36].

Finally, developing lightweight models that generalize well across tasks is essential, addressing trade-offs between performance, complexity, and robustness in multimodal systems [7]. Future research directions also include scaling the CLIPSONIC method to larger video datasets and exploring tri-modal models that incorporate audio, vision, and language [62]. By pursuing these research opportunities, the field of multimodal learning can advance significantly, leading to the development of more sophisticated, adaptable, and efficient AI systems.

6 Conclusion

Multimodal learning has emerged as a pivotal advancement in artificial intelligence, effectively integrating varied data types such as text, images, audio, and video to overcome the constraints of unimodal systems. This integration enhances model performance and resilience, exemplified by methodologies like the Multimodal Learning with Alternating (MLA) approach, which addresses modality laziness and boosts performance across diverse datasets. Dual-Modality Prompt Tuning (DPT) further adapts vision-language models for specific tasks, utilizing class-aware visual prompts.

In speech recognition, incorporating global context within the ASR pipeline has led to notable performance improvements, indicating promising directions for future research. The BIKE framework's success in video recognition tasks underscores the potential of bidirectional knowledge exploration from pre-trained vision-language models. Similarly, the II-CLVM framework excels in video-music retrieval, overcoming challenges related to noisy data and achieving superior results with minimal training samples.

The enhancement of audio representation learning through multimodal techniques is evident in the establishment of new state-of-the-art benchmarks. The HighMMT model demonstrates significant improvements in high-modality scenarios, achieving strong results across various tasks. In medical imaging, adapting CLIP models enhances performance by improving image-text alignment.

Frameworks like MXM-CLR have shown superiority in cross-modal representation learning, outperforming existing methods. The COTS model achieves exceptional results in image-text and video-text retrieval tasks, improving both retrieval performance and inference efficiency. The MFM approach maintains robustness to missing modalities while learning effective multimodal representations, and the ATD method demonstrates enhanced predictive modeling across benchmarks.

Future research should focus on refining training strategies to manage multimodal data complexities, integrating innovative techniques such as generative models, and improving representation of underrepresented content. Understanding the alignment between data modalities and technical designs could foster a cohesive framework for multimodal machine learning. The integration of knowledge graphs into vision-language models presents substantial performance improvements, meriting further exploration to address current limitations. Advancements in model architectures, data integration strategies, and ethical considerations will be essential to fully realize the potential of multimodal learning.

References

[1] Jabeen Summaira, Xi Li, Amin Muhammad Shoib, and Jabbar Abdul. A review on methods and applications in multimodal deep learning, 2022.

[2] Jiahao Qin. Zoom and shift are all you need, 2024.

[3] Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion*, 81:203–239, 2022.

[4] Maria Lymperaiou and Giorgos Stamou. A survey on knowledge-enhanced multimodal learning, 2024.

[5] David M. Chan, Shalini Ghosh, Debmalya Chakrabarty, and Björn Hoffmeister. Multi-modal pre-training for automated speech recognition, 2022.

[6] Luyu Wang, Pauline Luc, Adria Recasens, Jean-Baptiste Alayrac, and Aaron van den Oord. Multimodal self-supervised learning of general audio representations, 2021.

[7] Paul Pu Liang. Foundations of multisensory artificial intelligence, 2024.

[8] Ye Zhu, Yu Wu, Nicu Sebe, and Yan Yan. Vision+x: A survey on multimodal learning in the light of data, 2024.

[9] Jean Park, Kuk Jin Jang, Basam Alasaly, Sriharsha Mopidevi, Andrew Zolensky, Eric Eaton, Insup Lee, and Kevin Johnson. Assessing modality bias in video question answering benchmarks with multimodal large language models, 2024.

[10] Wei Chen, Weiping Wang, Li Liu, and Michael S. Lew. New ideas and trends in deep multimodal content understanding: A review, 2020.

[11] Songtao Li and Hao Tang. Multimodal alignment and fusion: A survey. *arXiv preprint arXiv:2411.17040*, 2024.

[12] Jabeen Summaira, Xi Li, Amin Muhammad Shoib, Songyuan Li, and Jabbar Abdul. Recent advances and trends in multimodal deep learning: A review, 2021.

[13] Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, et al. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *arXiv preprint arXiv:2408.01319*, 2024.

[14] Zhou Lu. On the computational benefit of multimodal learning, 2023.

[15] Clayton Cohn, Eduardo Davalos, Caleb Vatral, Joyce Horn Fonteles, Hanchen David Wang, Meiyi Ma, and Gautam Biswas. Multimodal methods for analyzing learning and training environments: A systematic literature review, 2024.

[16] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations, 2019.

[17] Yinghui Xing, Qirui Wu, De Cheng, Shizhou Zhang, Guoqiang Liang, Peng Wang, and Yanning Zhang. Dual modality prompt tuning for vision-language pre-trained model, 2023.

[18] Youze Wang, Wenbo Hu, Yinpeng Dong, Hanwang Zhang, Hang Su, and Richang Hong. Exploring transferability of multimodal adversarial samples for vision-language pre-training models with contrastive learning, 2025.

[19] Tristan Karch, Yoann Lemesle, Romain Laroche, Clément Moulin-Frier, and Pierre-Yves Oudeyer. Contrastive multimodal learning for emergence of graphical sensory-motor communication, 2023.

[20] Luyu Wang and Aaron van den Oord. Multi-format contrastive learning of audio representations, 2021.

[21] Bin Shan, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil 2.0: Multi-view contrastive learning for image-text pre-training, 2022.

[22] Maxwell Aladago, Lorenzo Torresani, and Soroush Vosoughi. Semantic compositions enhance vision-language contrastive learning, 2024.

[23] Thomas M. Hehn, Julian F. P. Kooij, and Dariu M. Gavrila. How do cross-view and cross-modal alignment affect representations in contrastive learning?, 2022.

[24] Xinhao Mei, Xubo Liu, Jianyuan Sun, Mark D. Plumbley, and Wenwu Wang. On metric learning for audio-text cross-modal retrieval, 2022.

[25] Bin Zhao, Maoguo Gong, and Xuelong Li. Audiovisual video summarization, 2021.

[26] Ishaan Singh Rawal, Alexander Matyasko, Shantanu Jaiswal, Basura Fernando, and Cheston Tan. Dissecting multimodality in videoqa transformer models by impairing modality fusion, 2024.

[27] Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. Multimodality representation learning: A survey on evolution, pretraining and its applications, 2024.

[28] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, Louis-Philippe Morency, and Ruslan Salakhutdinov. High-modality multimodal transformer: Quantifying modality interaction heterogeneity for high-modality representation learning, 2023.

[29] Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models, 2023.

[30] Pranav Arora et al. Domain-agnostic multi-modal video retrieval. Master's thesis, 2023.

[31] Haocheng Dai and Sarang Joshi. Refining skewed perceptions in vision-language models through visual representations, 2025.

[32] Chang Che, Qunwei Lin, Xinyu Zhao, Jiaxin Huang, and Liqiang Yu. Enhancing multimodal understanding with clip-based image-to-text transformation, 2024.

[33] Hadi Hadizadeh, S. Faegheh Yeganli, Bahador Rashidi, and Ivan V. Bajić. Mutual information analysis in multimodal learning systems, 2024.

[34] Mahdi M. Kalayeh, Shervin Ardeshir, Lingyi Liu, Nagendra Kamath, and Ashok Chandrashekar. On negative sampling for audio-visual contrastive learning from movies, 2022.

[35] Feilong Chen, Xiuyi Chen, Shuang Xu, and Bo Xu. Improving cross-modal understanding in visual dialog via contrastive learning, 2022.

[36] Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment, 2025.

[37] Gustavo Adolfo Vargas Hakim, David Osowiechi, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Moslem Yazdanpanah, Ismail Ben Ayed, and Christian Desrosiers. Clipartt: Adaptation of clip to new domains at test time, 2024.

[38] Abdelhamid Haouhat, Slimane Bellaouar, Attia Nehar, and Hadda Cherroun. Modality influence in multimodal machine learning, 2023.

[39] Donghuo Zeng, Jianming Wu, Gen Hattori, Yi Yu, and Rong Xu. Learning explicit and implicit latent common spaces for audio-visual cross-modal retrieval, 2021.

[40] Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval, 2022.

[41] Yangming Zhou, Qichao Ying, Zhenxing Qian, Sheng Li, and Xinpeng Zhang. Multimodal fake news detection via clip-guided learning, 2022.

[42] Di Hu, Feiping Nie, and Xuelong Li. Dense multimodal fusion for hierarchically joint representation, 2018.

[43] Yaowei Li, Ruijie Quan, Linchao Zhu, and Yi Yang. Efficient multimodal fusion via interactive prompting, 2023.

[44] Wenqi Zhu, Jiale Cao, Jin Xie, Shuangming Yang, and Yanwei Pang. Clip-vis: Adapting clip for open-vocabulary video instance segmentation, 2024.

[45] Delong Chen, Zhao Wu, Fan Liu, Zaiquan Yang, Huaxi Huang, Ying Tan, and Erjin Zhou. Protoclip: Prototypical contrastive language image pretraining, 2023.

[46] Chaoya Jiang, Wei Ye, Haiyang Xu, Miang yan, Shikun Zhang, Jie Zhang, and Fei Huang. Vision language pre-training by contrastive learning with cross-modal similarity regulation, 2023.

[47] Zhixiong Zeng and Wenji Mao. A comprehensive empirical study of vision-language pre-trained model for supervised cross-modal retrieval, 2022.

[48] Jorge Sánchez and Rodrigo Laguna. Cross-modal coordination across a diverse set of input modalities, 2024.

[49] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning, 2023.

[50] Fatema Hasan, Yulong Li, James Foulds, Shimei Pan, and Bishwaranjan Bhattacharjee. Teach me with a whisper: Enhancing large language models for analyzing spoken transcripts using speech embeddings, 2023.

[51] Tiantian Feng, Daniel Yang, Digbalay Bose, and Shrikanth Narayanan. Can text-to-image model assist multi-modal learning for visual recognition with visual modality missing?, 2024.

[52] Ye Wang, Bowei Jiang, Changqing Zou, and Rui Ma. Mxm-clr: A unified framework for contrastive learning of multifold cross-modal representations, 2023.

[53] Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, and Dinggang Shen. Clip in medical imaging: A comprehensive survey, 2024.

[54] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval, 2021.

[55] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text, 2022.

[56] Yogesh Kumar and Pekka Marttinen. Improving medical multi-modal contrastive learning with expert annotations, 2024.

[57] Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does clip benefit visual question answering in the medical domain as much as it does in the general domain?, 2021.

[58] Tongkun Su, Jun Li, Xi Zhang, Haibo Jin, Hao Chen, Qiong Wang, Faqin Lv, Baoliang Zhao, and Yin Hu. Design as desired: Utilizing visual question answering for multimodal pre-training, 2024.

[59] Benoit Dufumier, Javiera Castillo-Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning?, 2024.

[60] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip, 2022.

[61] Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. Deep multimodal learning for audio-visual speech recognition, 2015.

[62] Hao-Wen Dong, Xiaoyu Liu, Jordi Pons, Gautam Bhattacharya, Santiago Pascual, Joan Serrà, Taylor Berg-Kirkpatrick, and Julian McAuley. Clipsonic: Text-to-audio synthesis with unlabeled videos and pretrained language-vision models, 2023.

[63] Zeyu Chen, Pengfei Zhang, Kai Ye, Wei Dong, Xin Feng, and Yana Zhang. Start from video-music retrieval: An inter-intra modal loss for cross modal retrieval, 2024.

[64] Gaurav Verma, Rohit Mujumdar, Zijie J. Wang, Munmun De Choudhury, and Srijan Kumar. Overcoming language disparity in online content classification with multimodal learning, 2022.

[65] Xianbing Zhao, Soujanya Poria, Xuejiao Li, Yixin Chen, and Buzhou Tang. Toward robust multimodal learning using multimodal foundational models, 2024.

[66] Ziyi Yang, Yuwei Fang, Chenguang Zhu, Reid Pryzant, Dongdong Chen, Yu Shi, Yichong Xu, Yao Qian, Mei Gao, Yi-Ling Chen, Liyang Lu, Yujia Xie, Robert Gmyr, Noel Codella, Naoyuki Kanda, Bin Xiao, Lu Yuan, Takuya Yoshioka, Michael Zeng, and Xuedong Huang. i-code: An integrative and composable multimodal learning framework, 2022.

[67] Yao-Hung Hubert Tsai, Martin Q. Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multimodal routing: Improving local and global interpretability of multimodal language analysis, 2020.

[68] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang. Audio-visual speech enhancement using multimodal deep convolutional neural networks, 2022.

[69] Jingyou Xie, Jiayi Kuang, Zhenzhou Lin, Jiarui Ouyang, Zishuo Zhao, and Ying Shen. Flex-clip: Feature-level generation network enhanced clip for x-shot cross-modal retrieval, 2024.

[70] Guoli Song, Shuhui Wang, Qingming Huang, and Qi Tian. Harmonized multimodal learning with gaussian process latent variable models, 2019.

[71] William Theisen and Walter Scheirer. N-modal contrastive losses with applications to social media data in trimodal space, 2024.

[72] Alex Andonian, Shixing Chen, and Raffay Hamid. Robust cross-modal representation learning with progressive self-distillation, 2022.

[73] Haoxuan You, Luowei Zhou, Bin Xiao, Noel Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. Learning visual representation from modality-shared contrastive language-image pre-training, 2022.

[74] Guiming Cao, Kaize Shi, Hong Fu, Huaiwen Zhang, and Guandong Xu. Aple: Token-wise adaptive for multi-modal prompt learning, 2024.

[75] Paul Lerner, Olivier Ferret, and Camille Guinaudeau. Cross-modal retrieval for knowledge-based visual question answering, 2024.

[76] Paul Pu Liang, Peter Wu, Liu Ziyin, Louis-Philippe Morency, and Ruslan Salakhutdinov. Cross-modal generalization: Learning in low resource modalities via meta-alignment, 2020.

[77] Normalized contrastive learning for text-video retrieval.

[78] Samrajya Thapa, Koushik Howlader, Subhankar Bhattacharjee, and Wei le. More: Multi-modal contrastive pre-training with transformers on x-rays, ecgs, and diagnostic report, 2024.

[79] Yuxuan Ding, Lingqiao Liu, Chunna Tian, Jingyuan Yang, and Haoxuan Ding. Don't stop learning: Towards continual learning for the clip model, 2022.

[80] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context, 2024.

[81] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts, 2022.

[82] Wei Chow, Juncheng Li, Qifan Yu, Kaihang Pan, Hao Fei, Zhiqi Ge, Shuai Yang, Siliang Tang, Hanwang Zhang, and Qianru Sun. Unified generative and discriminative training for multi-modal large language models. *arXiv preprint arXiv:2411.00304*, 2024.

[83] Maciej Pawłowski, Anna Wróblewska, and Sylwia Sysko-Romańczuk. Does a technique for building multimodal representation matter? – comparative analysis, 2022.

[84] Abdul Aziz A. B and A. B Abdul Rahim. Topological perspectives on optimal multimodal embedding spaces, 2024.

[85] Yunwei Ren and Yuanzhi Li. On the importance of contrastive loss in multimodal learning, 2023.

[86] Paul Pu Liang, Chun Kai Ling, Yun Cheng, Alex Obolenskiy, Yudong Liu, Rohan Pandey, Alex Wilf, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal learning without labeled multimodal data: Guarantees and applications, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.