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, enabling the integration of diverse data types—text, images, audio, and video—to enhance understanding and prediction across various domains [1]. This approach effectively addresses the heterogeneity challenge that traditional methods, which typically convert data into a common vector space, struggle to overcome [2]. By harnessing the complementary strengths of different sensory modalities, MML overcomes the limitations of unimodal systems, leading to more comprehensive and accurate machine learning models [3].

Recent surveys emphasize the integration of external knowledge sources into vision-and-language (VL) representation learning, highlighting MML's potential to bridge knowledge gaps in VL tasks, particularly in scenarios with limited labeled multimodal data, where understanding modality interactions enhances reasoning and task performance [4]. In automated speech recognition (ASR), the combination of audio and visual data has proven to enhance performance, especially in noisy environments [5]. This integration is vital for developing robust systems that function effectively under challenging conditions, as incorporating video information significantly improves audio representation learning for various audio-related tasks [6].

The theoretical and computational foundations of MML are evolving, with ongoing research identifying common themes and open questions in the field [7]. This evolution drives the development of sophisticated multimodal systems capable of addressing real-world complexities. The integration of multiple modalities not only enhances the performance and interpretability of machine learning sys-

Generated by SurveyX.cn.

www.SurveyX.cn

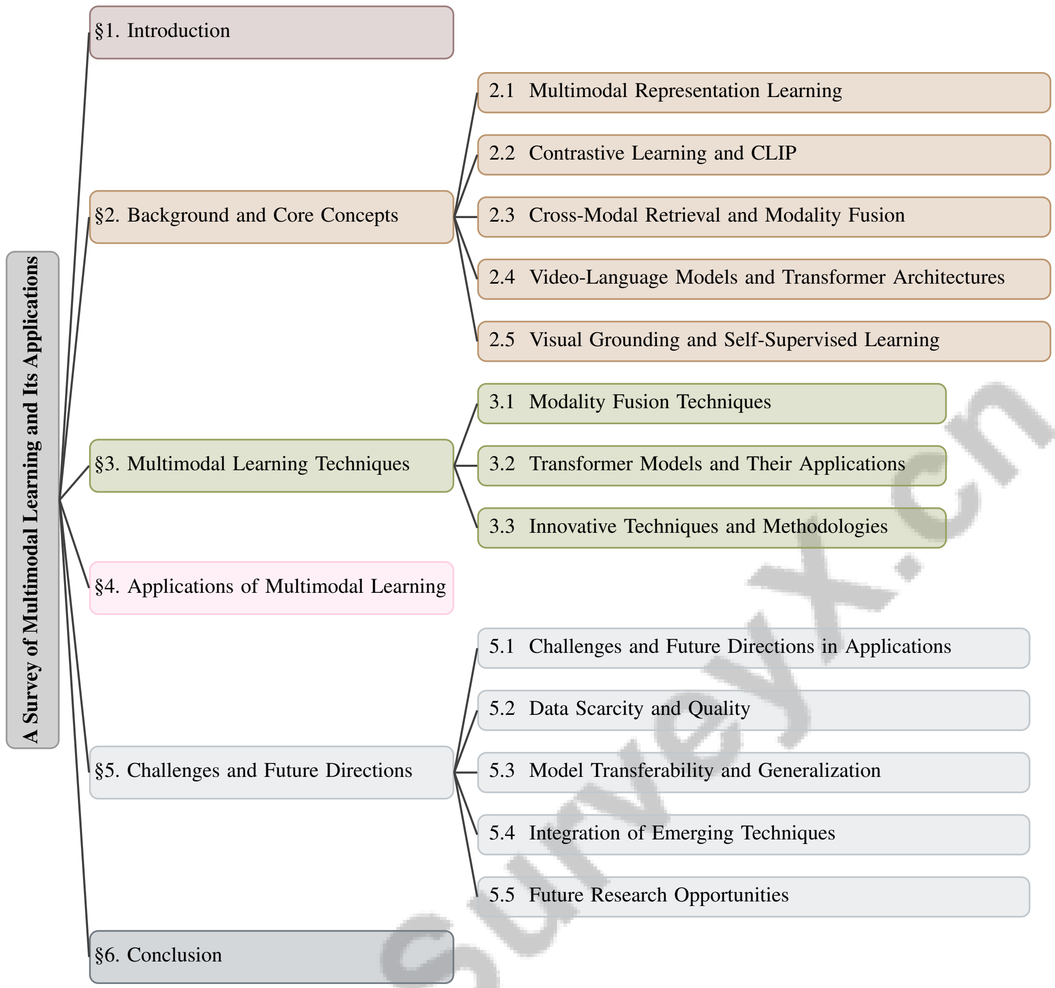


Figure 1: chapter structure

tems but also broadens their applicability across diverse domains, establishing MML as a cornerstone of modern artificial intelligence research [8].

1.2 Motivation for the Survey

This survey aims to address 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, the complexity of integrating multiple data types presents formidable challenges, necessitating innovative methodologies [3]. The survey explores these advancements, focusing on modality heterogeneity, connections, and interactions, which are crucial for understanding the foundational elements of multimodal machine learning [7].

A critical impetus for this survey is the need to address modality bias in existing video question answering benchmarks, which limits the assessment of models' ability 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. Additionally, it aims to fill 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 in vision-and-language (VL) models, where limited comprehension of commonsense, factual, and temporal knowledge can hinder real-world performance [4].

The survey delivers a comprehensive analysis of the current multimodal learning landscape, focusing on integrating diverse data types—including text, images, audio, and video. It addresses co-learning challenges such as handling missing or noisy modalities and disparities in feature representation.

www.SurveyX.cn

AI-generated, for reference only.

Furthermore, it highlights recent advancements in multimodal alignment and fusion techniques, providing strategic insights for effective integration across applications like social media analysis, medical imaging, and visiolinguistic tasks. By systematically reviewing over 200 relevant studies, this work aims to guide future research toward enhancing the scalability, robustness, and generalizability of multimodal learning systems [3, 10, 11, 4].

1.3 Relevance in Current Research Landscape

The multimodal learning landscape has transformed significantly, with recent research highlighting its critical role in enhancing model performance across applications such as emotion recognition and video captioning [12]. This progress is driven by the development of Multimodal Large Language Models (MLLMs), which have shown substantial improvements in processing and integrating multimodal data, reinforcing their relevance amid ongoing research and technological advancements [13].

Understanding the theoretical underpinnings of multimodal learning is essential for navigating the complexities of integrating diverse data types. Recent studies emphasize that a solid grasp of these foundations is vital for advancing research in this domain [14]. MML plays a pivotal role in bridging semantic and heterogeneity gaps, especially between image and text data, which is crucial for seamless modality integration [15].

In educational contexts, the integration and analysis of multimodal data—including speech, video, and eye gaze—have enhanced our understanding of learner behaviors and outcomes. This integration is instrumental in developing educational technologies that respond more effectively to learners' nuanced needs [11]. Moreover, significant progress in modeling multimodal interactions has improved performance in critical applications such as healthcare and affective computing [7].

The significance of multimodal learning is underscored by its potential to advance the field toward 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 performing complex tasks with greater accuracy and efficiency [8]. As research evolves, integrating multimodal learning techniques remains a cornerstone in the pursuit of advanced artificial intelligence systems.

1.4 Structure of the Survey

This survey is structured to provide a comprehensive exploration of multimodal learning, beginning with an introduction that underscores the significance and motivation behind this field. The introduction discusses the relevance of multimodal learning in the current research landscape, highlighting its transformative impact across various domains. The subsequent "Background and Core Concepts" section examines key theories and methodologies, 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 methodologies employed in the field, particularly focusing on self-supervised learning and modality fusion techniques alongside innovative approaches propelling the field forward. The survey then transitions into "Applications of Multimodal Learning," exploring practical implementations 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 to 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.

AI-generated, for reference only.

www.SurveyX.cn

2 Background and Core Concepts

2.1 Multimodal Representation Learning

Multimodal representation learning integrates text, images, audio, and video into a unified framework, enhancing task performance by leveraging each modality's strengths and overcoming unimodal limitations [3]. Key challenges include learning complex intra-modal and cross-modal interactions and ensuring robustness against missing or noisy modalities [16]. Techniques like AV-BERT, which employs masked language modeling for audio-visual data integration, significantly improve speech recognition in challenging environments [5]. Dual-modality prompt tuning (DPT) further enhances multimodal learning by adjusting text and visual prompts concurrently [17].

Current research categorizes multimodal co-learning challenges and methods, offering a comprehensive overview of effective integration strategies [3]. Investigations into adversarial robustness in Vision-Language Pre-training (VLP) models underscore the need for a unified framework to generate adversarial text-image pairs, crucial for developing robust systems [18]. Contrastive learning methods, such as CURVES, bridge the semantic gap between modalities by generating graphical utterances that denote visual referents [19]. Benchmarks for discriminative audio representation learning address the integration of raw audio with spectral representations [20].

2.2 Contrastive Learning and CLIP

Contrastive learning enhances multimodal integration by leveraging similarities and differences across data modalities, ensuring balanced representation and addressing modality competition [21]. The Contrastive Language–Image Pretraining (CLIP) model exemplifies this approach, achieving notable results in image-text retrieval and zero-shot classification using large datasets of image-text pairs [22]. Challenges such as false negatives and improper minimization of mutual information can hinder optimization, leading to lost semantic information [23, 19].

Innovative methods like CLIPC generate composite image-caption pairs, enhancing vision-language model robustness [22]. The Multi-View Contrastive learning framework models intra-modal and inter-modal correlations, overcoming single-view limitations in many VLP models [21]. Theoretical foundations emphasize the importance of accurately estimating mutual information in high-dimensional spaces to understand modality relationships [23]. Focusing on positive and negative pairs improves training dynamics, enhancing alignment and balance in contrastive learning models [19].

2.3 Cross-Modal Retrieval and Modality Fusion

Cross-modal retrieval and modality fusion are vital for bridging semantic and distributional gaps between modalities like text, images, audio, and video. Cross-modal retrieval uses queries from one modality to retrieve semantically relevant information from another, addressing challenges posed by inconsistent data distributions [24]. The COTS model exemplifies effective cross-modal interactions, enhancing image-text retrieval efficiency [22]. However, many approaches fail to utilize multifold observations, leading to suboptimal representations [19].

Modality fusion combines diverse data types into unified representations, leveraging each modality's strengths to mitigate inefficiencies from independent processing, crucial for optimal performance when certain modalities are missing or poorly represented [24]. Aligning and integrating modalities during training and testing necessitates advanced strategies for robust model performance [23]. Despite these challenges, modality fusion is essential for robust systems, enabling alignment and integration of diverse modalities to enhance AI performance [22].

2.4 Video-Language Models and Transformer Architectures

Video-language models and transformer architectures are pivotal in multimodal learning, particularly in understanding and generating video content through language. These models utilize both video and textual data for robust interpretations. For instance, the AudioVisual Recurrent Network (AVRN) enhances video summarization by integrating audio and visual features [25]. Methodologies include explicit and implicit alignment methods and fusion frameworks like kernel-based and attention-based approaches [10], crucial for tasks like Video Question Answering (VideoQA) [26].

AI-generated, for reference only.

Transformer architectures have shown exceptional effectiveness in multimodal settings, with surveys documenting their evolution and impact [27]. The High Modality Multimodal Transformer (HighMMT) employs cross-modal attention mechanisms, enhancing multimodal representations [28]. Approaches like the Bidirectional Cross-Modal Knowledge Exploration (BIKE) enhance video recognition capabilities through Video-to-Text and Text-to-Video knowledge transfer [29]. The Domain-Agnostic Multi-Modal Video Retrieval (DAMMVR) method integrates audio and video features, showcasing transformers' adaptability in complex retrieval tasks [30].

2.5 Visual Grounding and Self-Supervised Learning

Visual grounding and self-supervised learning are crucial for advancing multimodal frameworks by improving alignment and integration of diverse data modalities. Visual grounding associates linguistic descriptions with specific visual elements, enhancing context-aware AI systems and refining task accuracy [31]. Multi-layered architectures with differential learning rates have improved image-to-text transformation processes [32]. Advanced entropy estimation techniques, like those in InfoMeter, emphasize mutual information's role in visual grounding [33].

Self-supervised learning exploits data's intrinsic structure to generate supervisory signals, enabling models to learn useful representations without extensive labeled datasets. This approach is advantageous in multimodal settings where labeled data are scarce. Techniques improve understanding of long-form content by integrating audio and visual information [34]. In automatic speech recognition, masked language modeling within self-supervised frameworks enhances performance [5].

The synergy between visual grounding and self-supervised learning is exemplified by contrastive learning methods that elucidate relationships between visual and textual elements [35]. Volume-based alignment measures, instead of traditional cosine similarity, facilitate geometric alignment across modalities, enriching semantic information [36]. Methods like CLIPArTT, which use automatic text prompt construction to generate pseudo-labels, improve test-time adaptation, enhancing multimodal systems' robustness [37].

3 Multimodal Learning Techniques

Category	Feature	Method
Modality Fusion Techniques	Prompt and Attention Strategies	PMF[38], DPT[17]
	Layer and Network Fusion	DMF[39]
	Adversarial and Robustness Approaches	Pri3D[23], VLP-attack[18]
	Latent Space and Interaction Techniques	AVCMR[40], ATD[2]
Transformer Models and Their Applications	Open-Vocabulary and Prototype Learning	ProtoCLIP[41], CLIP-VIS[42]
	Knowledge Transfer and Enhancement	ALKD[43]
	Cross-Modal and Multimodal Retrieval	COTS[44], CLIP4CMR[45]
	Semantic Optimization and Alignment	SRCL[46], GRAM[36]
Innovative Techniques and Methodologies	Data Augmentation Techniques	GTI-MM[47], CLIPC[22]
	Multimodal Enhancement Strategies	MXM-CLR[48], C4C[49], EV2[21], FND-CLIP[50]

Table 1: This table provides a comprehensive overview of the latest advancements in multimodal learning techniques, categorized into Modality Fusion Techniques, Transformer Models and Their Applications, and Innovative Techniques and Methodologies. It details various methods and frameworks that enhance the integration and interaction of diverse data modalities, highlighting their significance in improving AI system capabilities across different domains.

The integration of diverse data modalities is pivotal for developing robust AI systems in multimodal learning, enhancing tasks like automated audio captioning and image-to-text transformation. This integration fosters nuanced understanding of complex interrelationships between diverse data types, facilitating improved feature alignment and semantic coherence across modalities [32, 15, 51, 52]. As illustrated in Figure 2, the hierarchical structure of multimodal learning techniques is categorized into three primary areas: Modality Fusion Techniques, Transformer Models and Their Applications, and Innovative Techniques and Methodologies. Each of these categories delves into specific techniques, models, and their applications, showcasing the integration and interaction of diverse data modalities to enhance AI systems' capabilities across various fields. Modality fusion techniques are integral to this exploration, forming the foundation for synthesizing diverse data forms and enabling complex interactions in multimodal learning. Table 4 presents a detailed classification of multimodal learning techniques, illustrating the integration of diverse data modalities to enhance AI systems.

AI-generated, for reference only.

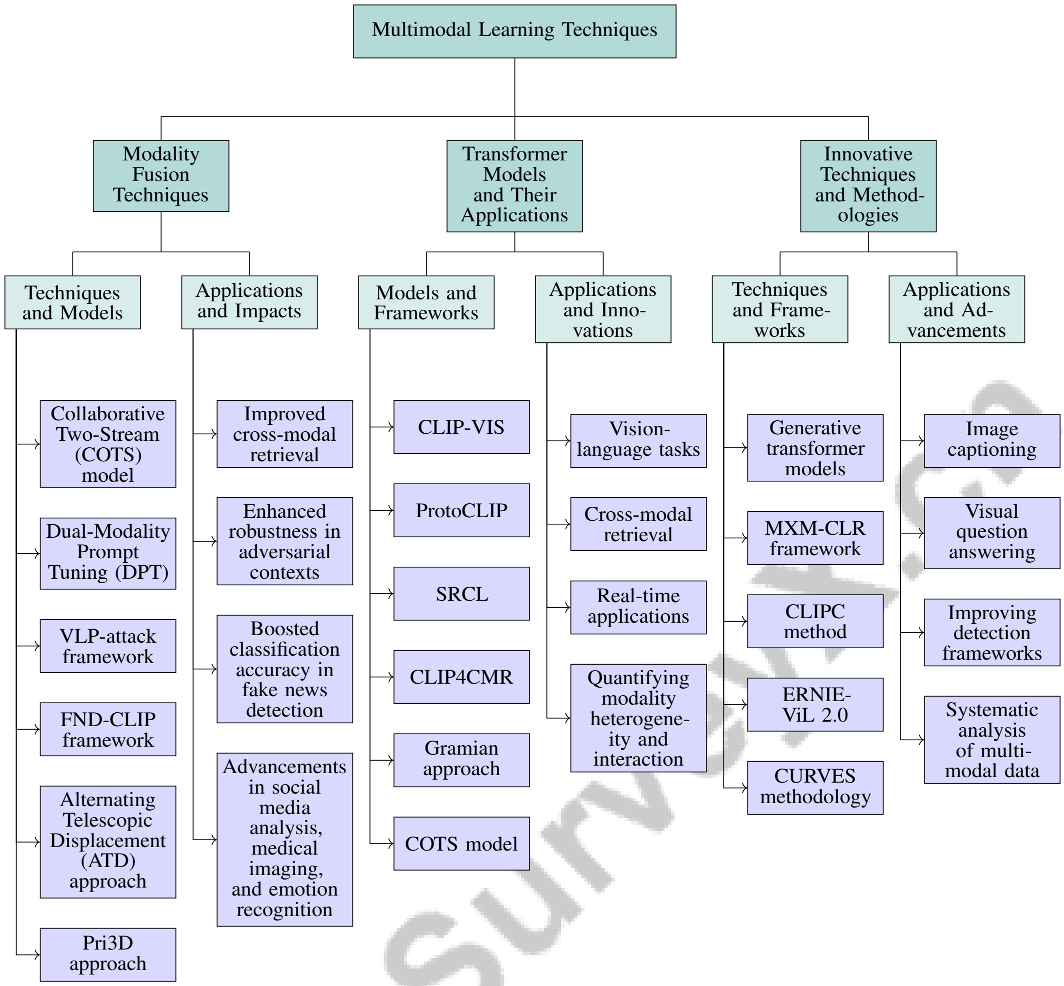


Figure 2: This figure illustrates the hierarchical structure of multimodal learning techniques, highlighting three primary categories: Modality Fusion Techniques, Transformer Models and Their Applications, and Innovative Techniques and Methodologies. Each category delves into specific techniques, models, and their applications, showcasing the integration and interaction of diverse data modalities to enhance AI systems' capabilities across various fields.

3.1 Modality Fusion Techniques

Method Name	Integration Techniques	Cross-Modal Interaction	Robustness and Accuracy
AVCMR[40]	Fusion Techniques	Shared Latent Subspaces	Maintain Semantic Integrity
COTS[44]	Adaptive Momentum Filter	Multiple Levels Interactions	Accurate Joint Representations
DPT[17]	Cross Attention	Class-aware Visual	Classification Accuracy
VLP-attack[18]	Contrastive Learning	Joint Information Exchange	Transferable Adversarial Samples
FND-CLIP[50]	Attention Mechanism	Cross-modal Similarity	Superior Performance
ATD[2]	Transformer-based Architecture	Alternating Transformations	Challenging Alignment Cases
Pri3D[23]	Contrastive Learning Approach	Cross-modal Contrastive Losses	Robust Visual Representations
PMF[38]	Interactive Prompting	Mutual Interactions	Memory-efficient Method
DMF[39]	Stacking Shared Layers	Multiple Learning Paths	Lower Training Loss

Table 2: Overview of various modality fusion techniques used in multimodal learning, highlighting their integration approaches, cross-modal interactions, and performance in terms of robustness and accuracy. The table summarizes methods such as AVCMR, COTS, DPT, and others, detailing their unique contributions to improving multimodal data processing and interpretation.

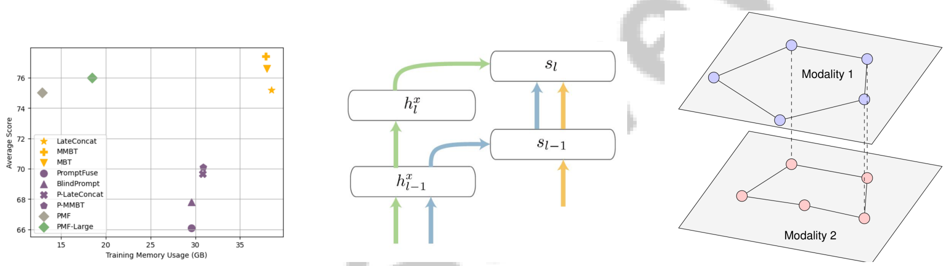
Modality fusion techniques are crucial for integrating diverse data types—audio, text, and video—enhancing model robustness and performance by creating unified representations that leverage each modality's strengths. Techniques using explicit and implicit latent subspaces preserve key

AI-generated, for reference only.

semantic information, improving cross-modal retrieval by acknowledging each modality's unique characteristics [40]. The Collaborative Two-Stream (COTS) model exemplifies advanced modality fusion through instance-level interaction via momentum contrastive learning and token-level interaction with masked vision-language modeling, facilitating efficient integration [44]. Dual-Modality Prompt Tuning (DPT) further enhances performance by generating class-aware visual prompts alongside text prompts, improving visual and textual data fusion [17].

In adversarial contexts, the VLP-attack framework integrates adversarial text and image generation into a single modality fusion technique, enhancing robustness [18]. The FND-CLIP framework employs the pretrained CLIP model to align features from text and images, using a modality-wise attention mechanism to boost classification accuracy in multimodal fake news detection [50]. Integrating different audio formats, akin to image processing augmentations, enhances learning, underscoring the significance of diverse modality representations in fusion [20]. The Alternating Telescopic Displacement (ATD) approach systematically shifts and expands feature representations across modalities, addressing alignment challenges [2].

The Pri3D approach, a self-supervised contrastive learning method utilizing RGB images and registered point clouds, underscores modality fusion's potential by learning visual representations without human annotations [23]. These techniques are integral to advancing multimodal learning, enabling AI systems to process and interpret complex information with greater accuracy and efficiency, enhancing capabilities across fields like social media analysis, medical imaging, and emotion recognition [53, 15, 10, 28, 51]. Table 2 provides a comprehensive summary of leading modality fusion techniques, illustrating their integration methods, cross-modal interactions, and effectiveness in enhancing robustness and accuracy in multimodal learning.



(a) Comparison of Average Scores and Training Memory Usage for Different Models[38]

(b) A neural network architecture with multiple layers and connections[39]

(c) Two modalities of data visualization[10]

Figure 3: Examples of Modality Fusion Techniques

As shown in Figure 3, effectively integrating information from diverse data sources is crucial for enhancing model performance and achieving robust results in multimodal learning. The first image illustrates a comparative analysis of average scores and training memory usage across different models, emphasizing the trade-offs between computational resources and performance outcomes. The second image delves into the architecture of a neural network with multiple layers and connections, essential for understanding how different modalities are processed and integrated. Lastly, the third image visualizes two distinct modalities represented by interconnected data points, demonstrating the relationships and interactions between modalities. Collectively, these images underscore the complexity and importance of modality fusion techniques in multimodal learning, providing insights into effectively integrating diverse data sources to enhance machine learning models [38, 39, 10].

3.2 Transformer Models and Their Applications

Transformer models are instrumental in advancing multimodal learning due to their robust ability to handle sequential data and capture complex dependencies across modalities. In vision-language tasks, models like CLIP-VIS adapt the frozen CLIP model for open-vocabulary video instance segmentation, highlighting transformers' flexibility in processing diverse data [42]. ProtoCLIP employs prototypes to guide representation learning in multimodal pretraining, enhancing representation grouping efficiency [41]. The SRCL method optimizes mutual information between image/text anchors and their negative counterparts, demonstrating transformers' potential in improving multimodal representation alignment [46].

AI-generated, for reference only.

The CLIP4CMR framework uses the CLIP model for supervised cross-modal retrieval, generating common representations for image and text data, bridging the semantic gap between modalities [45]. 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 COTS model exemplifies transformer efficiency in multimodal scenarios, achieving faster inference speeds compared to single-stream models while outperforming two-stream models in retrieval tasks [44]. This efficiency is crucial for real-time applications requiring rapid processing and integration of multimodal information. A loss function that accounts for multiple pairwise combinations of input modalities facilitates effective cross-modal representation learning, showcasing transformers' adaptability [54].

Transformers continue to enhance multimodal learning by enabling effective integration and interpretation of diverse data types. Through advanced learning techniques and innovative adaptations, transformers are central to multimodal research, facilitating the integration of diverse modalities—text, images, audio, and sensor data—allowing models like HighMMT and Meta-Transformer to learn and generalize efficiently across various applications. These models leverage novel metrics for quantifying modality heterogeneity and interaction, improving performance and efficiency in tasks like visual question answering and cross-modal retrieval. Consequently, transformers drive significant progress across numerous applications, from natural language processing to complex data mining tasks [55, 27, 15, 28].

As illustrated in Figure 4, the applications of transformer models span various domains, categorizing their use into vision-language tasks, cross-modal retrieval, and multimodal learning. This figure highlights key models such as CLIP-VIS, ProtoCLIP, SRCL, CLIP4CMR, GRAM, COTS, HighMMT, Meta-Transformer, and PCMC, showcasing the breadth and versatility of transformers in addressing complex AI challenges. The visual representation not only emphasizes the significance of these models but also reflects the ongoing evolution and integration of multimodal techniques in artificial intelligence research.

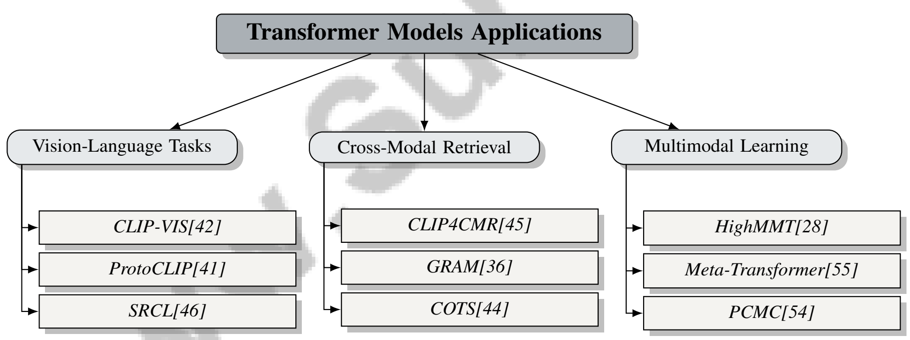


Figure 4: This figure illustrates the applications of transformer models in various domains. It categorizes the applications into vision-language tasks, cross-modal retrieval, and multimodal learning, highlighting key models such as CLIP-VIS, ProtoCLIP, SRCL, CLIP4CMR, GRAM, COTS, HighMMT, Meta-Transformer, and PCMC.

3.3 Innovative Techniques and Methodologies

Recent advancements in multimodal learning have introduced innovative techniques and methodologies that significantly enhance the integration and interaction of diverse data modalities. Generative transformer models, like the text-to-image model, enrich existing frameworks by generating new, meaningful data representations [47]. The MXM-CLR framework, with its multifold-aware hybrid loss (MFH), emphasizes considering multiple perspectives to enhance the robustness and accuracy of multimodal representations [48].

In cross-modal learning, the CLIPC method creates new training examples by blending images and merging captions, facilitating easier learning and improving downstream task performance [22]. The ERNIE-ViL 2.0 model constructs various visual/textual views, significantly enhancing cross-modal representation robustness and generalization [21]. The CURVES methodology facilitates the

AI-generated, for reference only.

www.SurveyX.cn

Method Name	Integration Techniques	Cross-Modal Learning	Application Domains
GTI-MM[47]	Synthetic Visual Data	Text-to-image Models	Visual Recognition
MXM-CLR[48]	Hybrid Loss	Positive Pairs	Cross-modal Retrieval
CLIPC[22]	Semantic Composition	Blending Images Captions	Medical Imaging
EV2[21]	Data Augmentation Techniques	Multi-view Contrastive	Cross-modal Retrieval
FND-CLIP[50]	Attention Mechanism	Cross-modal Similarity	Fake News Detection
C4C[49]	Pre-trained Representations	Temporal Features Incorporation	Video-text Retrieval

Table 3: This table presents a comparative overview of various multimodal learning methods, highlighting their integration techniques, cross-modal learning strategies, and application domains. The methods discussed include GTI-MM, MXM-CLR, CLIPC, EV2, FND-CLIP, and C4C, each contributing uniquely to advancements in fields such as visual recognition, cross-modal retrieval, medical imaging, and fake news detection.

emergence of a coherent language that generalizes across contexts, outperforming traditional methods reliant on discrete tokens [19].

In medical imaging, future research should focus on developing efficient training methods, exploring cross-modal applications, and enhancing CLIP models' generalizability [56]. The CLIP model's use in measuring cross-modal similarity and guiding feature fusion represents a significant advancement, improving accuracy and reliability in multimodal systems [50].

Innovative methodologies are driving significant advancements in multimodal learning, enhancing AI systems' ability to process and interpret complex multimodal information with increased accuracy and efficiency. By tackling fundamental challenges and employing innovative techniques, recent advancements in multimodal deep learning—such as integrating image and text data—are significantly advancing artificial intelligence. These developments enhance capabilities in tasks like image captioning and visual question answering while addressing critical issues like language disparity in online content classification. For instance, multimodal approaches have been shown to bridge performance gaps between English and non-English language models, improving detection frameworks for various applications, including crisis information and fake news. Furthermore, systematic analysis of multimodal data in learning environments reveals deeper insights into learner behaviors, while contrastive learning methods refine content representation in vision-language models, leading to improved generalization and robustness. Collectively, these efforts open new avenues for creating sophisticated AI systems [57, 58, 15, 11]. As shown in **??** and detailed in Table 3, the field of multimodal learning is rapidly evolving, with innovative techniques and methodologies emerging to enhance the integration and understanding of diverse data types. Table 3 provides a detailed comparison of recent multimodal learning methodologies, emphasizing their integration techniques, cross-modal learning approaches, and specific application domains. One innovative approach in video-text similarity calculation employs a Transformer encoder to bridge the gap between textual and visual inputs. This method utilizes distinct encoders for text and video, transforming sequences of words and frames into hidden state representations for similarity comparison, offering a robust framework for aligning video content with textual descriptions. In image-text question answering, deep learning models are being harnessed to tackle complex queries, designed to interpret and respond to questions based on visual content, with configurations tailored for both binary "Yes/No" questions and intricate multi-choice scenarios. Collectively, these examples underscore advanced machine learning frameworks' potential to facilitate seamless interaction and understanding across multiple modalities, paving the way for more sophisticated and intuitive AI systems [49, 15].

Feature	Modality Fusion Techniques	Transformer Models and Their Applications	Innovative Techniques and Methodologies
Integration Method	Unified Representations	Sequential Data Handling	Generative Transformers
Cross-Modal Interaction	Instance-level Interaction	Semantic Alignment	Blending Images/captions
Application Domain	Multimodal Learning	Vision-language Tasks	Medical Imaging

Table 4: This table provides a comprehensive comparison of multimodal learning techniques categorized into Modality Fusion Techniques, Transformer Models and Their Applications, and Innovative Techniques and Methodologies. It highlights key aspects such as integration methods, cross-modal interactions, and application domains, illustrating the diverse approaches employed to enhance AI systems through multimodal data integration.

AI-generated, for reference only.

4 Applications of Multimodal Learning

4.1 Healthcare and Medical Imaging

Multimodal learning has revolutionized healthcare and medical imaging by integrating various data types to improve diagnostic accuracy and patient outcomes. Techniques like contrastive learning, as seen in the MedCLIP model, enhance data efficiency and performance in medical image-text contrastive learning, forming a robust basis for medical diagnostics [59]. This method effectively aligns medical images with textual data, improving the interpretability and utility of multimodal datasets.

In medical imaging, CLIP-driven methods, particularly the eCLIP model, surpass traditional techniques in interpretability and performance, enhancing cross-modal retrieval by efficiently retrieving medical images based on textual queries [60]. Such advancements underscore the critical role of data integration in clinical settings where timely and accurate information retrieval is vital.

Multimodal Large Language Models (MLLMs) extend the capabilities of natural language processing and image understanding, enabling comprehensive analyses of Electronic Health Records (EHR) and improving healthcare solutions [13]. These models enhance patient data representation, supporting better healthcare delivery.

The PubMedCLIP model has shown promise in improving MedVQA systems, leading to more accurate medical image interpretation [61]. In tasks like Visual Question Answering (VQA), pre-training frameworks guide models to focus on specific pathological features, achieving superior performance in subsequent tasks like report generation and classification [62].

Empirical validations of multimodal learning frameworks, such as CoMM, across diverse healthcare datasets, demonstrate their applicability in real-world scenarios, highlighting the robustness and versatility of these techniques in addressing complex healthcare challenges [63].

4.2 Video and Audio Processing

Advancements in multimodal learning for video and audio processing have significantly improved retrieval, classification, and understanding tasks. The Wav2CLIP model exemplifies this progress by performing audio classification, retrieval, and captioning tasks, showcasing the potential of aligning audio and textual data to enhance retrieval accuracy [64].

In cross-modal retrieval, the GRACE method outperforms baseline methods on datasets like MSCOCO and Visual Genome, emphasizing the importance of generative capabilities in multimodal retrieval tasks [46]. The CoAVT model further illustrates the utility of multimodal learning in capturing complex audio-visual interactions [10].

Techniques like ModalityMirror have significantly improved audio classification performance, particularly where audio alone may be insufficient [65]. The LLM2CLIP framework enhances CLIP's performance across various tasks, demonstrating the utility of large language models in multimodal representation learning [66].

CMMixer provides a robust framework for cross-modal retrieval, capturing interdependencies between audio and video, leading to superior retrieval performance [67]. Fast adaptation of pretrained contrastive models has been validated on tasks like video question answering and text-to-video retrieval, underscoring the efficiency of contrastive learning techniques in multimodal contexts [6].

Experiments with the Clotho V2 and AudioCaps datasets illustrate the application of multimodal learning in audio processing [6]. Additionally, multimodal learning has been applied to video captioning, with experiments on the MSR-VTT and VATEX datasets comparing models using ImageNet Pre-training and CLIP, showcasing the enhanced capabilities of multimodal approaches in video processing [66].

The application of CLIP-powered models across diverse datasets, including audio, visual, and audio-visual questions, highlights the versatility and effectiveness of multimodal learning in addressing complex video and audio processing tasks [46]. These advancements underscore the transformative potential of multimodal learning in interpreting video and audio data across various applications.

AI-generated, for reference only.

4.3 Sentiment Analysis and Emotion Recognition

Multimodal learning has advanced sentiment analysis and emotion recognition by integrating diverse data types to enhance interpretability and accuracy. Techniques have been applied to sentiment analysis tasks, with language models evaluated on datasets designed for emotion detection, leveraging the complementary strengths of text, audio, and visual data to capture nuanced emotional and contextual information [57].

Frameworks like the Transformative Robust Multimodal Learning (TRML) have proven effective in processing incomplete data across modalities, enhancing the robustness of sentiment analysis systems [68]. This approach addresses challenges posed by missing or noisy data, ensuring high performance in sentiment analysis.

Innovative methodologies such as the Integrative and Composable Multimodal (iCODE) model have shown adaptability and effectiveness in sentiment and emotion analysis across various benchmarks [69]. These models enhance the accuracy of predictions by integrating multiple data sources, leading to a comprehensive understanding of emotional states.

Multimodal routing techniques have improved interpretability in sentiment analysis by effectively routing information between modalities, enabling better capture of emotional cues present in different data types [70]. In spoken language processing, models like Whisper have enhanced the understanding of emotional and contextual nuances, particularly in sentiment analysis tasks where integrating audio and textual data enriches emotional interpretation [43].

The integration of various modalities, including text and images, has demonstrated improvements in predictive accuracy and deeper insights into the relationships between different input types. Approaches like Multimodal Routing facilitate a nuanced understanding of individual modalities' contributions, enabling both global and local interpretability, which is crucial for addressing the black-box nature of existing models [70, 57, 71, 15, 11].

4.4 Multimodal Learning in Real-World Scenarios

Multimodal learning shows significant promise in real-world applications by integrating diverse data types to enhance performance across domains. The AVDCNN model exemplifies this potential by improving speech enhancement outcomes through audio-visual data integration, demonstrating the practical advantages of multimodal learning in challenging acoustic environments [72].

In cross-modal retrieval tasks, FLEX-CLIP outperforms state-of-the-art methods in zero-shot and few-shot scenarios, showcasing robust applications in real-world settings [73]. Harmonized multimodal Gaussian Process Latent Variable Models (GPLVM) have demonstrated superior performance in cross-modal retrieval tasks, validating the effectiveness of multimodal approaches in practical scenarios [74].

Trimodal models have advanced the understanding and classification of social media content, achieving state-of-the-art results in retrieval tasks involving diverse data types such as text, images, and videos [75]. The robustness of multimodal methods in adapting to natural distribution shifts indicates their potential for broader application in vision-language tasks [76].

Multimodal learning plays a pivotal role in generalized category discovery, achieving state-of-the-art results across datasets by leveraging integrated data types [77]. The MS-CLIP model highlights the benefits of parameter sharing in enhancing efficiency, particularly in zero-shot recognition and linear probing tasks [78].

The APLe framework demonstrates competitive generalization capabilities in scenarios with domain shifts, effectively addressing challenges related to prompt length and overfitting [79]. The GTI-MM model enhances data efficiency and robustness, particularly when visual modalities are missing, outperforming existing methods in various experimental setups [47]. This resilience in handling incomplete or noisy data is crucial in practical scenarios [3].

The practical applications of multimodal learning underscore its transformative potential in enhancing accuracy, robustness, and efficiency across diverse domains. By integrating multiple data modalities—such as text, images, audio, and video—these systems improve their capability to address intricate tasks and adapt to the dynamics of real-world environments. This integration enhances model accuracy by leveraging complementary information and facilitates knowledge transfer in scenarios

AI-generated, for reference only.

with limited data, revealing insights that individual modalities may overlook, thereby increasing the robustness and generalizability of machine learning applications across fields including social media analysis, medical imaging, and emotion recognition [10, 11].

5 Challenges and Future Directions

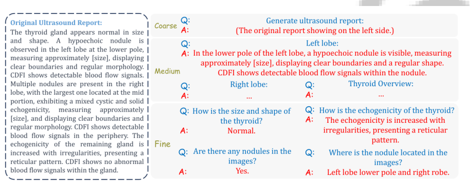
In examining the challenges and future directions of multimodal learning, it is crucial to identify the specific obstacles faced in applying these techniques. The complexities of integrating diverse data sources and ensuring model robustness are significant issues. The following subsections outline key challenges and propose future research directions to enhance the efficacy and adaptability of multimodal systems.

5.1 Challenges and Future Directions in Applications

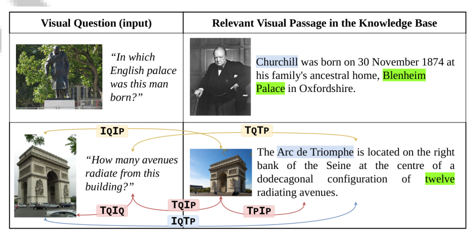
Multimodal learning applications encounter challenges due to missing modalities and high noise levels [3]. Current studies inadequately address these issues, highlighting the need for robust models. The discrete nature of text data complicates generating transferable multimodal adversarial samples, affecting adversarial robustness [18].

Future research should explore advanced fusion techniques, such as attention mechanisms, to enhance the granularity of multimodal representations [40]. Hybrid knowledge integration approaches could improve model generalizability across vision-language tasks [4]. Developing explainable fake news detection systems is critical for identifying suspicious elements in news articles [50]. Extending frameworks like MXM-CLR to video data could enhance multimodal learning systems' robustness [48].

In multimodal adversarial learning, generating transferable adversarial samples is crucial. The ATD method improves accuracy in multimodal tasks while reducing complexity [2]. Future research should explore cross-modal alignment in self-supervised learning [23]. Addressing these challenges can lead to more effective, scalable, and adaptable multimodal solutions.



(a) The image shows a table with three columns and five rows, each containing a question and an answer.[62]



(b) Visual Question Answering with Knowledge Base[80]

Figure 5: Examples of Challenges and Future Directions in Applications

Figure 5 illustrates challenges and future directions in VQA systems and their integration with knowledge bases. The first image shows a structured table titled "Original Ultrasound Report," enhancing understanding through detailed questions and answers. The second image demonstrates a VQA system using a knowledge base to improve information retrieval and comprehension through structured data, emphasizing ongoing challenges in refining these technologies for broader application [62, 80].

5.2 Data Scarcity and Quality

Data scarcity and quality significantly impact multimodal learning model development. Reliance on large labeled datasets, scarce in specialized domains like medical imaging, poses a challenge [56]. This scarcity is compounded by the need for multi-scale features and specialized knowledge.

Data quality is another critical issue, as models often overfit to noise, reducing generalization capabilities [67]. In contrastive learning, performance depends on the quality of constructed views,

AI-generated, for reference only.

with poor-quality views leading to suboptimal outcomes [21]. Benchmarks often fail to capture complexities in audio-text retrieval, highlighting the need for comprehensive evaluation frameworks [24].

Methods struggle with dissimilar modalities, particularly with missing or noisy inputs. The AV-BERT method exemplifies this challenge, as it relies on visual data availability during training [5]. Vision-language models' limited effectiveness in low-data contexts underscores the need for approaches leveraging data diversity [22].

Efforts to address these challenges include updating outdated benchmarks and selecting clusters for effective training [34]. Improving data availability and quality and developing methodologies that minimize reliance on large annotated datasets can advance robust and adaptable multimodal solutions.

5.3 Model Transferability and Generalization

Model transferability and generalization are pivotal challenges in multimodal learning. Discrepancies between input and output spaces across modalities complicate knowledge transfer, necessitating effective alignment of representation spaces [81]. This challenge is compounded by the dependency on training data quality and diversity [32].

The NCL framework illustrates difficulties in generalizing across different query distributions [82]. The MoRE framework highlights challenges in applying learned representations to new datasets, indicating adaptability limitations [83].

Continual learning algorithms face challenges in maintaining zero-shot capabilities while fine-tuning on new tasks [84]. The MoVA framework enhances generalization through adaptive expert selection, reducing bias from irrelevant vision encoders [85].

The DPT approach underscores the difficulty in modifying visual features extracted by image encoders, limiting downstream task applicability [17]. The COTS model faces computational challenges in calculating similarity scores for all query-candidate pairs, highlighting efficiency and interaction trade-offs [44].

Parameter sharing in multimodal models can enhance performance but may not fully capture unique modality characteristics [78]. Balanced modalities during training are critical, as imbalances can lead to suboptimal expert utilization [86].

5.4 Integration of Emerging Techniques

Integrating emerging techniques in multimodal learning is essential for addressing current challenges and enhancing AI system robustness. Promising directions include refining training strategies to accommodate multimodal data complexities. The Sugar framework shows potential in improving model performance across various contexts [87].

Incorporating novel methods like generative models, which create synthetic visual data, offers a transformative approach to enriching multimodal frameworks [47]. Hybrid loss functions that consider multiple positive observations enhance robustness by overcoming limitations of single-fold-oriented methods [48].

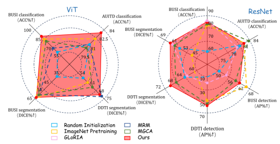
Attention mechanisms and hybrid knowledge integration improve model generalizability across diverse vision-language tasks [4]. Advanced entropy estimation techniques, such as InfoMeter, emphasize mutual information's role in improving visual grounding and accuracy [33].

Emerging methodologies, such as the ATD method, optimize feature alignment and reduce computational complexity, improving accuracy in multimodal tasks [2]. These advancements highlight the potential of optimized methodologies in advancing multimodal learning.

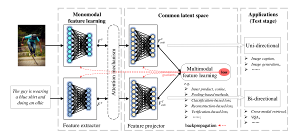
Integrating emerging techniques is vital for overcoming challenges and driving progress. Innovative techniques and optimized methodologies can significantly enhance multimodal systems' performance, enabling effective processing and interpretation of complex datasets. Comparative analyses of multimodal representation techniques underscore the importance of selecting appropriate approaches based on data characteristics and tasks. Leveraging multiple modalities boosts predictive accuracy and reveals intricate patterns often overlooked in single-modality analyses, providing a comprehensive understanding of underlying data [57, 71, 15, 88, 11].

AI-generated, for reference only.

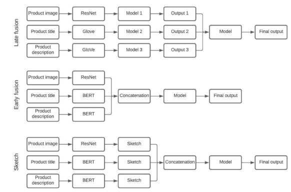
www.Survey/.cn



(a) Comparison of ViT and ResNet Models in Image Classification and Segmentation Tasks[62]



(b) Multimodal Feature Learning and Its Applications[15]



(c) Three Different Fusion Strategies for Product Image and Description Analysis[88]

Figure 6: Examples of Integration of Emerging Techniques

Figure 6 illustrates the integration of emerging techniques in AI and machine learning, presenting challenges and opportunities for future advancements. The examples highlight three key areas: the comparison of ViT and ResNet models in image classification and segmentation tasks, ongoing explorations of model architectures; multimodal feature learning, showcasing a two-stage process that enriches data representation; and three distinct fusion strategies for product image and description analysis—Late fusion, Early fusion, and Sketch—demonstrating innovative methods for synthesizing information from multiple modalities. These examples signal promising directions for future research [62, 15, 88].

5.5 Future Research Opportunities

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

Exploring the implications of local dimension on model performance and developing robust architectures that bridge modality gaps are crucial [89]. Applying theoretical insights from contrastive loss functions to complex models and investigating additional training techniques are promising directions [90]. Creating complementary questions that necessitate genuine integration of multiple modalities can further advance understanding of modality interactions [9].

Research should refine parameter grouping strategies and apply models like HighMMT to diverse modalities and tasks, enhancing adaptability and performance [28]. Improving robustness against dataset biases and exploring additional modalities for enhanced retrieval are vital areas for further exploration [30].

Future work could focus on improving interaction quantification accuracy and exploring applications in self-supervised learning, leading to more effective frameworks [91]. Investigating unsupervised or semi-supervised learning techniques can enhance multimodal approaches in scenarios with limited labeled data [16].

Refinements in negative sampling strategies and applying methods to other media forms offer additional avenues for advancement [34]. Further exploration of the GRAM method's application to diverse datasets and enhancing robustness against misalignment are essential [36].

Developing lightweight models that generalize well across tasks is crucial, addressing trade-offs between performance, complexity, and robustness [7]. Future research should consider scaling the CLIPSONIC method to larger video datasets and exploring tri-modal models incorporating audio, vision, and language [66]. Pursuing these research opportunities can significantly advance multimodal learning, leading to sophisticated, adaptable, and efficient AI systems that effectively leverage diverse data modalities.

www.SurveyX.cn

AI-generated, for reference only.

6 Conclusion

Multimodal learning has emerged as a pivotal advancement in artificial intelligence, offering a robust framework for integrating diverse data modalities such as text, images, audio, and video. This integration enhances model performance and robustness, surpassing the constraints of unimodal systems. Techniques like the Multimodal Learning with Alternating approach address challenges such as modality laziness, yielding improved outcomes across various datasets. Furthermore, advancements in Dual-Modality Prompt Tuning have significantly enhanced the adaptability of vision-language models, underscoring the effectiveness of class-aware visual prompts.

In speech recognition, incorporating global context into the ASR pipeline has led to notable performance improvements, opening new avenues for multimodal ASR research. The BIKE framework, through effective bidirectional knowledge exploration, achieves state-of-the-art results in video recognition tasks. Similarly, the II-CLVM framework excels in video-music retrieval by effectively managing noisy data and achieving superior results with limited training samples.

Recent advancements in audio representation learning have set new benchmarks, highlighting multimodal learning's potential. The HighMMT model demonstrates enhanced performance in high-modality scenarios, showcasing efficiency across various tasks. In medical imaging, adaptations of CLIP models have improved task performance, facilitating efficient image-text alignment.

Frameworks like MXM-CLR have significantly advanced cross-modal representation learning, outperforming existing methodologies. The COTS model has achieved remarkable retrieval performance and inference efficiency, setting new standards in image-text and video-text retrieval tasks. The MFM approach has demonstrated exceptional performance across multiple datasets, effectively learning robust multimodal representations. Additionally, the ATD method has shown its capability to integrate multimodal data, enhancing predictive modeling and achieving leading performance across benchmarks.

Future research should focus on refining training strategies to address the complexities of multimodal data and explore the integration of emerging techniques such as generative models. Improving the representation of underrepresented content within training datasets is essential. A deeper understanding of the alignment between data modalities and technical designs may foster a more unified framework for multimodal machine learning. Moreover, integrating knowledge graphs into vision-language models has shown substantial performance enhancements, warranting further exploration to overcome existing limitations. Progress in model architectures, data integration strategies, and ethical considerations will be crucial for fully realizing the potential of multimodal learning.

www.survey.cn

AI-generated, for reference only.

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

[11] 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.

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

[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.

AI-generated, for reference only.

www.SurveyX.cn

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

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

[40] 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.

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

AI-generated, for reference only.

www.SurveyX.cn

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

[43] 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.

[44] 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.

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

[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] 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.

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

[49] 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.

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

[51] Liwen Tan, Yin Cao, and Yi Zhou. Edtc: enhance depth of text comprehension in automated audio captioning, 2024.

[52] Abhishek Arora, Xinmei Yang, Shao-Yu Jheng, and Melissa Dell. Linking representations with multimodal contrastive learning, 2024.

[53] Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation, 2023.

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

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

[56] 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.

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

[58] Yichao Cai, Yuhang Liu, Zhen Zhang, and Javen Qinfeng Shi. Clap: Isolating content from style through contrastive learning with augmented prompts, 2024.

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

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

[61] 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.

[62] 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.

AI-generated, for reference only.

www.SurveyX.cn

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

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

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

[66] 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.

[67] 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.

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

[69] 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.

[70] 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.

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

[72] 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.

[73] 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.

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

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

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

[77] Enguang Wang, Zhimao Peng, Zhengyuan Xie, Fei Yang, Xialei Liu, and Ming-Ming Cheng. Unlocking the multi-modal potential of clip for generalized category discovery, 2024.

[78] 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.

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

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

[81] 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.

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

AI-generated, for reference only.

www.SurveyX.cn

[83] 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.

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

[85] 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.

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

[87] 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.

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

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

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

[91] 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.

www.SurveyX.cn

AI-generated, for reference only.

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.

www.SurveyX.cn

AI-generated, for reference only.