A Survey of Multimodal Large Language Models for Video Understanding and Analysis

www.surveyx.cn

Abstract

Multimodal Large Language Models (MLLMs) have emerged as pivotal tools in advancing video understanding and analysis by integrating diverse data modalities such as text, images, and video. This survey highlights the transformative potential of MLLMs, underscoring their superior performance in multimodal comprehension and vision perception tasks. Models like MiniGPT4-Video exemplify the enhanced capability of MLLMs to process complex video content, emphasizing their critical role in improving video comprehension. The survey identifies comprehensive evaluation methods as essential for the continued development of MLLMs, with benchmarks like CODIS highlighting areas for further enhancement in context-dependent visual comprehension. Despite significant advancements, challenges persist, particularly in capturing complex human-object interactions and temporal dynamics in videos. Current benchmarks reveal the need for refined ensemble mechanisms and innovative techniques to enhance model adaptability to evolving video formats. Future research should focus on integrating additional modalities and improving supervised retrievers for multimodal in-context learning. While current vision-based LLMs may not fully replicate human cognitive processes, their potential to revolutionize video understanding and analysis is significant. Ongoing research promises to address existing limitations, unlocking new possibilities for sophisticated and reliable video analysis applications. The proposed VideoLLM framework validates the efficacy of transferring LLM capabilities to video understanding tasks, reflecting the future potential of MLLMs in enhancing video comprehension and analysis.

1 Introduction

1.1 Significance of MLLMs in Video Understanding

Multimodal Large Language Models (MLLMs) have emerged as pivotal tools in video comprehension, addressing the pressing demand for advanced systems capable of managing and interpreting extensive video data generated by various applications [1]. By integrating multiple modalities—text, images, and video—MLLMs substantially enhance the accuracy and depth of video analysis.

A primary advantage of MLLMs lies in their ability to create rich textual representations from visual inputs, which is essential for improving classification accuracy and enriching the context for Video Question Answering (Video QA) systems, key in extracting meaningful insights from video content [2]. Furthermore, MLLMs overcome the limitations of traditional language models that focus primarily on text, enhancing tasks such as video quality assessment (VQA) that require robust algorithms to ensure optimal video quality in streaming media [3].

The development of MLLMs highlights the need to address unique challenges in video analysis, including increased computational demands and effective semantic alignment across modalities [4]. Additionally, the challenge of misinformation, particularly multimodal claims combining text, images, and videos, is effectively tackled by MLLMs, which are well-equipped to mitigate these issues [5].

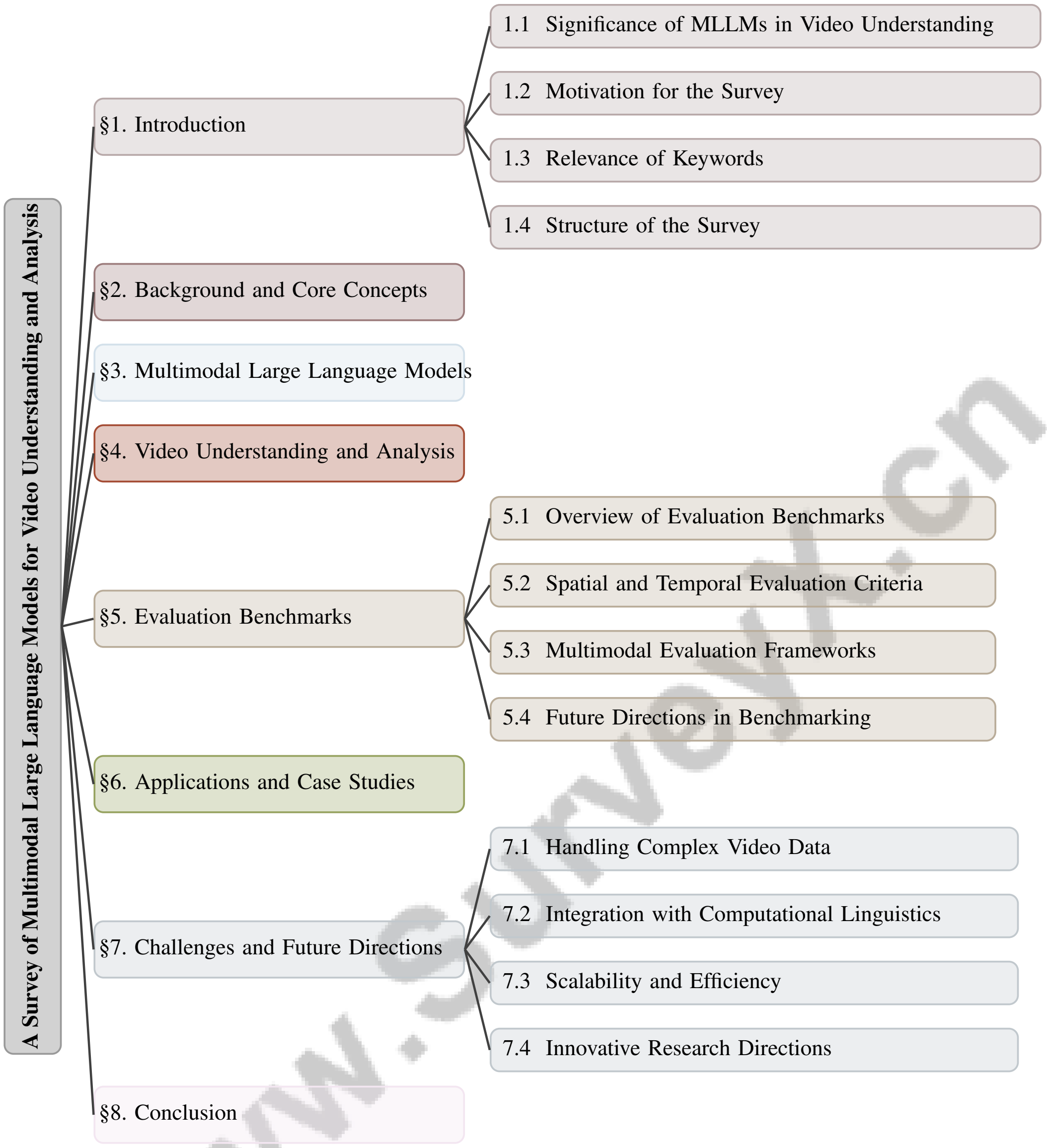

A Survey of Multimodal Large Language Models for Video Understanding and Analysis
§1. Introduction
1.1 Significance of MLLMs in Video Understanding
1.2 Motivation for the Survey
1.3 Relevance of Keywords
1.4 Structure of the Survey
§2. Background and Core Concepts
§3. Multimodal Large Language Models
§4. Video Understanding and Analysis
§5. Evaluation Benchmarks
5.1 Overview of Evaluation Benchmarks
5.2 Spatial and Temporal Evaluation Criteria
5.3 Multimodal Evaluation Frameworks
5.4 Future Directions in Benchmarking
§6. Applications and Case Studies
§7. Challenges and Future Directions
7.1 Handling Complex Video Data
7.2 Integration with Computational Linguistics
7.3 Scalability and Efficiency
7.4 Innovative Research Directions
§8. Conclusion

Figure 1: chapter structure

As MLLMs evolve, they are set to play a crucial role in bridging different modalities, advancing the field towards more integrated and intelligent video analysis systems [3]. This evolution underscores the transformative potential of MLLMs in enhancing video comprehension and analysis capabilities, paving the way for sophisticated and reliable applications [6]. For instance, MiniGPT4-Video exemplifies this potential by effectively processing both visual and textual data to enhance video comprehension and analysis [7]. Furthermore, the diverse data integration by MLLMs not only improves video comprehension but also parallels their applications in bioimage analysis, showcasing their versatility across various domains [8].

1.2 Motivation for the Survey

This survey is motivated by the urgent need to advance MLLMs in video understanding, as current models often struggle to capture the temporal dynamics inherent in video sequences [7]. Existing approaches are typically designed for specific tasks, lacking the flexibility required for comprehensive video analysis when integrated with language. This survey aims to address this gap by advocating for generalized models capable of tackling diverse video understanding challenges [1].

AI-generated, for reference only.

A significant focus of this survey is to confront the persistent issue of semantic misalignment in MLLMs, particularly when processing disparate visual inputs, which is critical for coherent video analysis [3]. The survey explores design choices that can enhance video-LMM performance, promoting systematic studies to optimize video understanding capabilities [4].

The rapid evolution of MLLMs necessitates a comprehensive evaluation of their visual comprehension abilities within broader contexts, highlighting the need for robust benchmarks to propel research in multimodal understanding [9]. This survey emphasizes augmenting visual knowledge integration in MLLMs, addressing the limitations of current methods that restrict image understanding [2].

Additionally, the survey tackles the semantic gap in multimodal data processing, which can lead to inaccurate outputs and societal risks. It aims to enhance collaborative inference between LLMs and other pre-trained models while managing computational efficiency [6]. The complexity of narrative videos necessitates flexible multimodal interactions, a challenge this survey seeks to address through context interleaving strategies [10].

Finally, this survey contributes to the field by providing insights into the potential of MLLMs in assisting fact-checkers and improving misinformation detection accuracy, thereby underscoring the broader societal implications of these technologies [5]. Through these efforts, the survey aims to deliver a comprehensive understanding of MLLMs, identify key areas for enhancement, and propose future research directions to bolster the integration and efficacy of multimodal AI systems in video understanding and analysis.

1.3 Relevance of Keywords

The selected keywords are crucial for understanding the dynamics of MLLMs and their application in video analysis. "Multimodal Large Language Models" encapsulates the integration of textual and visual data to enhance video comprehension, a concept central to frameworks like GPT4Video, which illustrates how LLMs, in conjunction with visual feature extractors and generative models, can advance cohesive video understanding and generation [11].

"Video Understanding" and "Video Analysis" represent the core objectives of MLLMs, emphasizing the integration of large language models into video comprehension tasks that remain underexplored, thus highlighting the importance of multimodal approaches for comprehensive video analysis [12].

"Evaluation benchmarks" are vital for assessing MLLM performance in video analysis, ensuring that these models meet rigorous standards across diverse video types and data modalities. Comprehensive benchmarks like Video-MME, MLLM-Bench, and MMBench-Video address the need for high-quality assessments by incorporating extensive datasets, expert annotations, and innovative evaluation methodologies. These benchmarks facilitate the evaluation of MLLMs in real-world scenarios, enhancing our understanding of their capabilities in processing sequential visual data and driving advancements in effective video analysis [13, 14, 15].

"Multimodal AI and Computational Linguistics" are interdisciplinary fields crucial for the development of MLLMs, which integrate various data types—text, images, and audio—to enhance capabilities in complex tasks like natural language understanding and visual processing, thereby advancing AI applications across multiple domains [16, 17, 18, 19, 20]. Multimodal AI enhances the integration of diverse data types, improving MLLMs' ability to process complex video inputs, while computational linguistics provides the necessary frameworks for interpreting and generating textual data, thus amplifying the overall efficacy of MLLMs in video analysis.

The integration of these keywords establishes a foundation for a deeper exploration of the transformative potential of MLLMs in video understanding and analysis. This exploration is particularly vital as it underscores the relevance of MLLMs in enhancing AI systems' capabilities to process and interpret complex multimodal data, addressing existing gaps in evaluating video composition understanding, and facilitating comprehensive multimodal context fusion. Innovations such as the VidComposition benchmark and the Video-MME evaluation framework emphasize the necessity for rigorous assessments of MLLMs' performance across diverse video contexts, ultimately paving the way for advancements that could bridge the performance gap between human and machine understanding in sequential visual data [21, 22, 14, 23].

AI-generated, for reference only.

1.4 Structure of the Survey

This survey is structured into several sections, each addressing critical aspects of MLLMs and their application in video understanding and analysis. Section 1 introduces the topic, emphasizing the significance of MLLMs and the motivation behind this survey, while also discussing key terms and providing a roadmap for the paper. Section 2 delves into the background and core concepts, offering definitions and exploring the interdisciplinary nature of multimodal AI and the evolution of MLLMs.

Section 3 focuses on the development and architecture of MLLMs, detailing design and training techniques and explaining how they integrate multiple data modalities for enhanced analysis. Section 4 examines the role of MLLMs in video understanding and analysis, with subsections dedicated to video comprehension, temporal reasoning, and the challenges of analyzing dynamic video content.

Section 5 evaluates the benchmarks used to assess MLLM performance in video analysis tasks, discussing spatial and temporal evaluation criteria, multimodal evaluation frameworks, and future directions in benchmarking. Section 6 provides an in-depth examination of applications and case studies that highlight the transformative impact of MLLMs across various fields, including healthcare, where they enhance clinical decision-making and patient engagement; education, through improved learning tools and resources; and narrative video analysis, illustrated by the VidComposition benchmark, which evaluates MLLMs' ability to analyze complex video compositions, revealing significant gaps in their understanding of intricate visual narratives [21, 22, 24, 20].

In Section 7, we explore significant challenges and prospective developments for MLLMs in video understanding, examining critical issues such as the complexities of processing intricate video data, integrating computational linguistics to enhance multimodal comprehension, and the need for improved scalability and efficiency in model training and application. We highlight the limitations of current benchmarks, such as VidComposition and Video-MME, which reveal substantial performance gaps in MLLMs' ability to analyze video compositions and their contextual dynamics. Additionally, we discuss innovative approaches like RED-VILLM and T2Vid that aim to optimize resource efficiency and enhance training methodologies, paving the way for more robust and effective video understanding systems [21, 14, 25, 23]. The survey concludes with Section 8, summarizing key findings and reflecting on the potential of MLLMs to advance video understanding and analysis.The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Definitions and Key Terms

Multimodal Large Language Models (MLLMs) are sophisticated AI systems that integrate text, images, and videos to enhance video understanding and analysis [2, 7]. A key feature is their ability to achieve effective semantic alignment across diverse data types, which is crucial for task-agnostic video comprehension and adaptability to various tasks without retraining [2]. MLLMs facilitate complex any-to-any understanding and generation across modalities, enhancing reasoning abilities by synthesizing information from diverse sources for real-world applications. Advances in training methodologies are improving contextual understanding and multimodal interactions [22, 24, 17, 26]. This capability is particularly beneficial for applications like multimodal convolutional neural networks (MM-CNNs) that generate natural language descriptions from video clips by leveraging both audio and visual data.

MLLMs are also being explored for analyzing complex biological images, showcasing their potential to integrate multimodal information for intricate analytical tasks [8]. This underscores the importance of ongoing research and development, as these models promise significant advancements in video analysis and understanding across a range of applications.

2.2 Interdisciplinary Nature of Multimodal AI

Multimodal AI integrates computational linguistics, visual cognition, and media analysis to enhance video analysis systems [5]. This interdisciplinary approach is essential for addressing complex cognitive tasks, advancing the understanding and interpretation of intricate video data. The convergence of vision and language in MLLMs exemplifies this integration, facilitating both video content analysis and generation [8]. Applications in clinical decision support and medical imaging benefit from this integration, enhancing diagnostic outcomes by combining visual and textual information.

Vision-language models improve classification tasks by integrating visual features with textual descriptions, demonstrating the transformative potential of combining these modalities. This is evident in audio-visual models and instruction tuning frameworks classified by modality types and application domains, ranging from video summarization and visual question answering to specialized fields like cooking and medical procedures. Leveraging MLLMs and innovative tuning strategies, these frameworks address technical challenges and optimize performance for diverse real-world tasks [27, 28, 22, 29].

The development of foundational models that integrate understanding and generation across modalities illustrates the potential for cohesive and contextually relevant video analysis. Recent surveys categorize research into stages like dataset construction, MLLM architecture design, and model fine-tuning, emphasizing the multifaceted nature of this field [8]. This interdisciplinary approach enhances video analysis by integrating techniques from natural language processing and computer vision, enabling sophisticated models capable of understanding, generating, and processing diverse data types, including text, audio, and visual content. This convergence fosters richer cross-modal understanding, exemplified by MLLMs' capabilities in tasks such as image-text generation and visual question answering, while addressing interpretability and explainability challenges crucial for high-stakes applications [30, 31, 18]. This approach lays the groundwork for more sophisticated AI systems, as demonstrated by frameworks focusing on multimodal reasoning through the combination of visual understanding and language processing.

2.3 Evolution of Multimodal Large Language Models

The evolution of MLLMs has significantly advanced the integration of text, images, audio, video, and physiological signals, enhancing AI systems' capabilities in understanding and analyzing complex real-world scenarios. These models enable richer cross-modal interactions and reasoning, addressing interpretability and transparency challenges crucial for high-stakes environments. As research progresses, MLLMs are expected to push AI boundaries, potentially leading toward artificial general intelligence [16, 32, 17, 18, 19]. This development can be delineated into phases, each contributing to the sophistication and efficacy of these models.

Initially, the focus was on single modality approaches, laying the groundwork for advancements in modality conversion and fusion techniques [33]. The transition from single modality to modality conversion (2000-2010) and fusion (2010-2020) marked a pivotal shift toward integrated systems capable of processing multiple data types concurrently. Recently, large-scale multimodal models (2020-present) have revolutionized the field, exemplified by MiniGPT4-Video, which integrates visual tokens from multiple frames with textual information, enhancing video understanding [7]. This advancement underscores MLLMs' potential in addressing complex tasks requiring nuanced interpretation of multimodal data.

Despite these advancements, challenges remain in optimizing the semantic alignment of visual and textual modalities. Models like the Semantic Alignment for Multimodal large language models (SAM) have been developed to improve visual and textual information integration [3]. Comprehensive benchmarks, such as InfiniBench, support the evaluation of MLLMs' comprehension and reasoning abilities over extended video content, highlighting the need for models capable of efficiently handling very long videos [34]. Furthermore, recent studies have explored MLLMs' potential in Image Quality Assessment (IQA), though this area remains less explored compared to other applications [35]. As MLLMs continue to evolve, enhancing their efficiency and effectiveness in processing and understanding multimodal data remains crucial. The continuous refinement of these models emphasizes the importance of innovation and interdisciplinary collaboration in advancing MLLMs' capabilities in video understanding and analysis.

In recent years, the development of Multimodal Large Language Models (MLLMs) has garnered significant attention in the field of artificial intelligence. These models represent a convergence of various modalities, including text, image, and sound, thereby enhancing their applicability and effectiveness in diverse tasks. To better understand this complex landscape, we can refer to Figure 2, which illustrates the hierarchical structure of MLLMs. This figure encompasses key aspects such as development and architecture, model design and training techniques, and multimodal integration. It highlights innovations, specific model examples, and evaluation components that are critical in the various stages of development and training. Furthermore, the figure delineates innovative strategies and benchmarks in model design, as well as advancements and future directions in multimodal integration. By examining this structured overview, we gain valuable insights into the intricate interplay of these elements, which collectively define the current state and future potential of MLLMs.

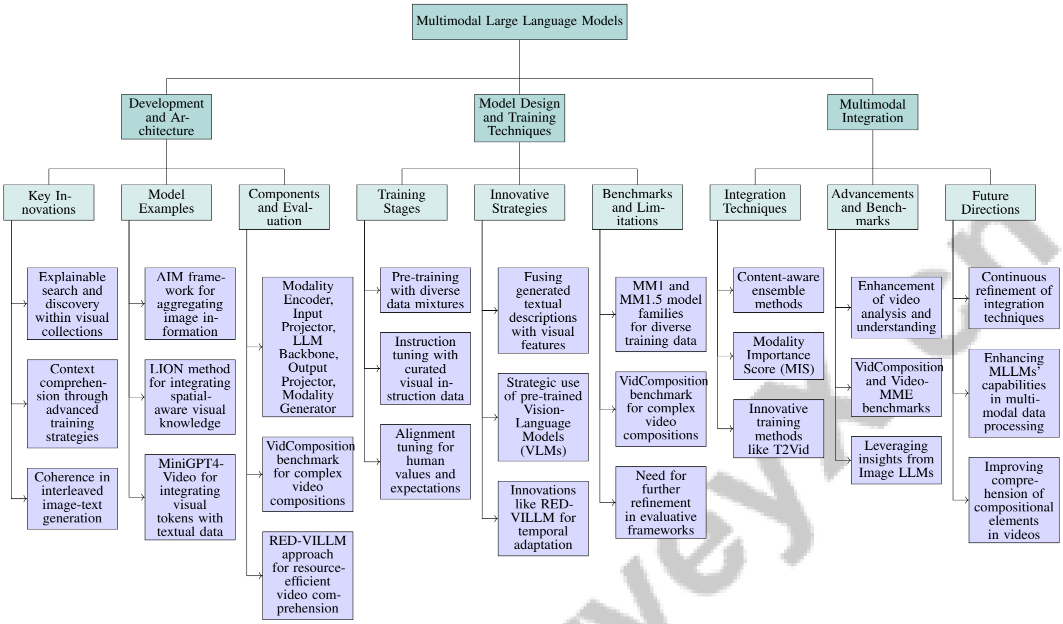

Multimodal Large Language Models
Development and Architecture
Model Design and Training Techniques
Multimodal Integration
Key Innovations
Model Examples
Components and Evaluation
Training Stages
Innovative Strategies
Benchmarks and Limitations
Integration Techniques
Advancements and Benchmarks
Future Directions
Explainable search and discovery within visual collections
Context comprehension through advanced training strategies
Coherence in interleaved image-text generation
AIM framework for aggregating image information
LION method for integrating spatial-aware visual knowledge
MiniGPT4-Video for integrating visual tokens with textual data
Modality Encoder, Input Projector, LLM Backbone, Output Projector, Modality Generator
VidComposition benchmark for complex video compositions
RED-VILLM approach for resource-efficient video comprehension
Pre-training with diverse data mixtures
Instruction tuning with curated visual instruction data
Alignment tuning for human values and expectations
Fusing generated textual descriptions with visual features
Strategic use of pre-trained Vision-Language Models (VLMs)
Innovations like RED-VILLM for temporal adaptation
MM1 and MM1.5 model families for diverse training data
VidComposition benchmark for complex video compositions
Need for further refinement in evaluative frameworks
Content-aware ensemble methods
Modality Importance Score (MIS)
Innovative training methods like T2Vid
Enhancement of video analysis and understanding
VidComposition and Video-MME benchmarks
Leveraging insights from Image LLMs
Continuous refinement of integration techniques
Enhancing MLLMs' capabilities in multi-modal data processing
Improving comprehension of compositional elements in videos

Figure 2: This figure illustrates the hierarchical structure of Multimodal Large Language Models (MLLMs), encompassing development and architecture, model design and training techniques, and multimodal integration. It highlights key innovations, model examples, and evaluation components in development, training stages, innovative strategies, and benchmarks in design, as well as integration techniques, advancements, and future directions in multimodal integration.

3 Multimodal Large Language Models

3.1 Development and Architecture

Method Name	Architectural Components	Integration Strategies	Evaluation Frameworks
AIM[36]	Fused Virtual Tokens	Aggregates Image Information	Cider, Vqa Accuracy
LION[2]	Modality Encoder	Stage-wise Instruction-tuning	Evaluation Method
M4V[7]	Modality Encoder	Visual Tokens	Multiple Benchmarks

Table 1: Comparison of architectural components, integration strategies, and evaluation frameworks of various Multimodal Large Language Models (MLLMs). The table provides an overview of the AIM, LION, and M4V methods, highlighting their unique approaches to integrating and evaluating multimodal data.

The architecture of Multimodal Large Language Models (MLLMs) is pivotal in advancing video understanding by integrating text, images, and video. Key innovations facilitate explainable search and discovery within visual collections, enhance context comprehension through advanced training strategies, and ensure coherence in interleaved image-text generation, addressing challenges of interpretability and transparency [37, 22, 38, 18]. The AIM framework exemplifies this by aggregating image information into textual labels' latent space, enhancing in-context learning [36]. The LION method further advances architecture by integrating spatial-aware visual knowledge with semantic evidence [2].

MiniGPT4-Video showcases the integration of visual tokens with textual data, enhancing video comprehension [7]. Models such as LLaVA-v1.6 and GPT-4V undergo continuous refinement for

AI-generated, for reference only.

improved video analysis [35]. MLLM architecture comprises five main components: Modality Encoder, Input Projector, LLM Backbone, Output Projector, and Modality Generator, each crucial for processing multimodal data [8]. These components ensure flexibility across various video analysis scenarios.

The development of models like MiniGPT4-Video and LION underscores efforts to optimize multimodal data processing, revealing their transformative potential in video understanding. The VidComposition benchmark evaluates MLLMs' ability to interpret complex video compositions, including camera movement and narrative structure. Integrating video foundation models with large language models, as seen in the RED-VILLM approach, offers resource-efficient strategies leveraging Image LLMs to enhance video comprehension, highlighting pathways for effective multimodal understanding [21, 23].

As illustrated in Figure 3, which depicts the development and architecture of MLLMs, several key innovations are highlighted, including explainable search interfaces, the AIM framework for efficient in-context learning, and the LION method for enhanced visual knowledge integration. Model examples such as MiniGPT4-Video, LLaVA-v1.6, and GPT-4V demonstrate advancements in multimodal data processing. Additionally, evaluation benchmarks like VidComposition, MLLM-IQA, and MM1 methods provide frameworks for assessing model performance and guiding future research. The "Data Processing" figure emphasizes transforming raw data into curated datasets, highlighting quality and safety in data management. The "Comparison of Model Performance on Different QA Tasks" figure illustrates model capabilities in handling distinct question-answering tasks, indicating variability in efficacy. The "Model Ablations and Data Ablations in Image-Text Generation" figure provides insights into model and data ablation processes, contributing to the refinement of image-text generation models. These figures encapsulate a multifaceted approach to developing robust multimodal language models, emphasizing data processing, performance evaluation, and model optimization [39, 40, 41]. Additionally, Table 1 presents a comparative analysis of different Multimodal Large Language Models (MLLMs), detailing their architectural components, integration strategies, and evaluation frameworks.

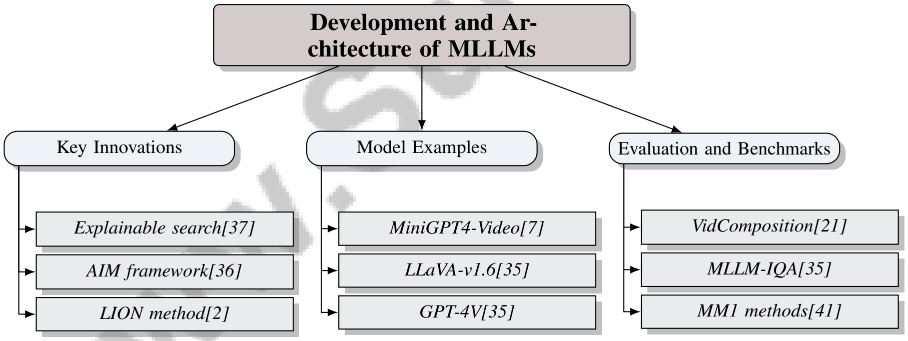


Figure 3: This figure illustrates the development and architecture of Multimodal Large Language Models (MLLMs), highlighting key innovations, model examples, and evaluation benchmarks. Key innovations include explainable search interfaces, the AIM framework for efficient in-context learning, and the LION method for enhanced visual knowledge integration. Model examples such as MiniGPT4-Video, LLaVA-v1.6, and GPT-4V demonstrate advancements in multimodal data processing. Evaluation benchmarks like VidComposition, MLLM-IQA, and MM1 methods provide frameworks for assessing model performance and guiding future research.

3.2 Model Design and Training Techniques

The evolution of Multimodal Large Language Models (MLLMs) in design and training techniques emphasizes integrating diverse data modalities for enhanced comprehension and generation. Stage-wise instruction-tuning strategies, such as the LION framework, integrate visual knowledge through innovative training methods [2]. MiniGPT4-Video's approach of concatenating visual tokens and subtitles for comprehensive inputs improves comprehension and analysis [7]. JourneyBench's human-machine-in-the-loop framework introduces challenging benchmarks for robust model development [42].

Training MLLMs involves pre-training, instruction tuning, and alignment tuning. Pre-training utilizes diverse data mixtures, including OCR data and synthetic captions, to establish a robust foundational model. Instruction tuning optimizes the model's ability to follow user instructions through curated visual instruction data, while alignment tuning refines responses to align with human values and expectations. These stages enhance MLLMs' capabilities in complex tasks like text-rich image understanding and multi-image reasoning [41, 15, 43, 44]. Techniques like fusing generated textual descriptions with visual features for zero-shot classification exemplify efforts to improve model adaptability across tasks.

Strategic use of pre-trained Vision-Language Models (VLMs) to summarize visual representations for language models exemplifies innovative strategies in MLLM training, ensuring efficient feature extraction and integration of visual features. Recent advancements in design and training methodologies for MLLMs underscore the critical role of modular architectures, dynamic integration mechanisms, and strategic training approaches in enhancing video understanding and analysis capabilities. Innovations like RED-VILLM transition from Image LLMs to Video LLMs, incorporating temporal adaptation within image fusion modules to extend comprehension to temporal elements with minimal resources. The MM1 and MM1.5 model families emphasize careful data curation and diverse training data mixtures to optimize performance across multimodal benchmarks. The VidComposition benchmark highlights MLLMs' limitations in analyzing complex video compositions, emphasizing the need for further refinement in evaluative frameworks and training methodologies [21, 41, 43, 23].

3.3 Multimodal Integration

Integrating multiple data modalities in Multimodal Large Language Models (MLLMs) is crucial for enhancing video analysis and understanding. Sophisticated techniques enable MLLMs to process and synthesize information from diverse sources, including text, images, and video. Content-aware ensemble methods, combining predictions from MLLMs with existing learning-based Blind Video Quality Assessment (BVQA) models, improve accuracy and reliability in video quality assessments [45].

The Modality Importance Score (MIS) is a novel metric for evaluating each modality's contribution to answering questions, enhancing understanding of modality integration and prioritization within MLLMs. By quantifying each modality's importance, MIS facilitates nuanced integration processes that optimize model performance across tasks [46].

These advancements underscore the transformative potential of effective multimodal integration in MLLMs, significantly enhancing their ability to conduct comprehensive and accurate video analysis. Benchmarks like VidComposition and Video-MME reveal current limitations in MLLMs' understanding of complex video compositions and sequential visual data. Leveraging insights from Image LLMs and innovative training methods such as T2Vid, researchers advance resource-efficient and scalable video understanding systems. This integration deepens comprehension of compositional elements and contextual dynamics in videos, improving MLLMs' overall performance across diverse multimodal tasks [21, 22, 14, 25, 23]. Continuous refinement of integration techniques and metrics like MIS is crucial for enhancing MLLMs' capabilities in processing and interpreting complex multimodal data.

4 Video Understanding and Analysis

In video understanding, integrating multiple modalities is crucial for comprehensively analyzing complex content. This section explores the key contributions of Multimodal Large Language Models (MLLMs) in enhancing video comprehension, particularly their ability to synthesize information from diverse sources. The following subsection details the specific role of MLLMs in video comprehension, illustrating how these models leverage multimodal data to deepen understanding of narrative structures and contextual elements within videos.

4.1 Role of MLLMs in Video Comprehension

MLLMs are pivotal in video comprehension through their integration and processing of text, images, and audio, facilitating comprehensive video data analysis. Recent studies emphasize the transformative impact of MLLMs, highlighting their ability to transfer reasoning capabilities from language models to video tasks [2]. The SAM model enhances semantic alignment and coherence in multimodal instructions, boosting MLLMs' video comprehension capabilities [3]. The AIM framework optimizes MLLMs for multimodal in-context learning, improving performance and efficiency [36]. Such advancements are vital for managing extensive input contexts needed for thorough video understanding, especially in tasks requiring synthesis of visual and textual information.

Textual integration significantly influences in-context learning performance in vision-language models, underscoring text's importance in video comprehension tasks [2]. The MovieSeq framework exemplifies the effectiveness of interleaved multimodal sequences in understanding narrative videos [10]. The TimeChat model incorporates visual and timestamp information, enhancing understanding of long videos and addressing temporal reasoning challenges [47]. This integration is particularly valuable for tasks necessitating nuanced interpretations of time-sensitive content. Despite advancements, MLLMs often fall short of human performance in context-dependent comprehension tasks [9]. However, ongoing developments, such as MiniGPT4-Video, show significant improvements in video question answering through effective visual and textual integration [7].

Advancements in MLLMs, driven by innovations in model design, integration strategies, and training techniques, underscore their transformative potential in enhancing video comprehension. MLLMs enhance performance by effectively integrating and processing complex multimodal data, such as text, images, and audio, showcasing their versatility in addressing real-world challenges [48, 22, 26, 17, 49].

4.2 Temporal Reasoning in Videos

Temporal reasoning is critical in video understanding, involving MLLMs' ability to interpret and reason about temporal dynamics and sequences of events. Capturing long-term dependencies and abstract temporal concepts remains challenging, often leading to fragmented temporal understanding [50]. Recent advancements introduce frameworks to address these challenges, such as the VideoAgent model, which uses a large language model to iteratively compile information for long-form video understanding, enhancing temporal sequence comprehension [51]. The TimeChat model improves temporal reasoning by associating visual content with timestamps, vital for tasks requiring accurate temporal localization and sequencing [47].

Benchmarks like MVBench emphasize evaluating temporal understanding in videos, particularly for tasks beyond single frames [52]. By focusing on temporal aspects, MVBench advances research by providing a structured framework for assessing MLLMs' ability to process and interpret temporal sequences. Integrating temporal reasoning capabilities into MLLMs is essential for enhancing video understanding, enabling models to process complex temporal dynamics effectively. Ongoing enhancements in MLLMs and evaluation frameworks highlight their potential to improve video content understanding over time. This evolution facilitates deeper comprehension of intricate video compositions—such as the interplay of visual elements, camera techniques, and narrative structures—while laying the groundwork for advanced video analysis applications. Benchmarks like VidComposition and Video-MME offer critical insights into MLLMs' capabilities in interpreting various video types and modalities, revealing performance gaps that indicate areas for further development. Frameworks like StreamingBench emphasize real-time processing capabilities, bridging the gap between human and model performance in streaming video contexts. These advancements suggest a future where MLLMs achieve more sophisticated video understanding, transforming video analysis and interpretation [21, 14, 25, 53, 23].

4.3 Challenges in Analyzing Dynamic Video Content

Analyzing dynamic video content with MLLMs presents challenges due to the intricate temporal and spatial reasoning required. A primary difficulty is the differing context windows for visual and language tokens, complicating LMMs' application to long video tasks [54]. This often leads to inefficiencies in processing extended sequences, hindering coherence over prolonged durations. Current benchmarks inadequately address long video complexities, focusing primarily on shorter clips, highlighting the need for comprehensive benchmarks for long-form video content. JourneyBench underscores these inadequacies, with advanced models struggling compared to traditional benchmarks [42].

As illustrated in Figure 4, the primary challenges in analyzing dynamic video content can be categorized into three key areas: context window issues, computational demands, and multimodal integration. The context window problems underscore the difficulties in managing visual context and the limitations of existing benchmarks. High computational demands for processing long visual sequences pose significant obstacles, as token aggregation can lead to information decay, diminishing models' effectiveness in capturing intricate temporal cues [55]. Balancing visual tokens while preserving spatial and temporal cues remains challenging, as existing methods often overlook either temporal or spatial details, resulting in incomplete analysis [56].

Furthermore, the figure emphasizes the complexities of multimodal integration, which complicates alignment and necessitates improved training techniques to enhance performance [19]. Additionally, extensive annotated datasets for diverse scenarios in bioimage analysis and the difficulty of automating analysis without human expertise highlight broader challenges in dynamic video content analysis [8].

Challenges in effectively analyzing complex and dynamic video data highlight the need for comprehensive evaluation frameworks and innovative methodologies. Recent benchmarks, such as VidComposition and the Complex Video Reasoning and Robustness Evaluation Suite (CVRR-ES), reveal significant performance gaps in MLLMs, particularly in understanding intricate video compositions and reasoning over diverse real-world scenarios. Video-MME, the first comprehensive evaluation benchmark for MLLMs in video analysis, emphasizes incorporating varied video types, temporal dimensions, and multi-modal inputs to enhance model assessment. Addressing these gaps is critical for advancing MLLMs' capabilities and ensuring effectiveness in realistic applications, from robotics to AI-assisted medical procedures [21, 14, 57].

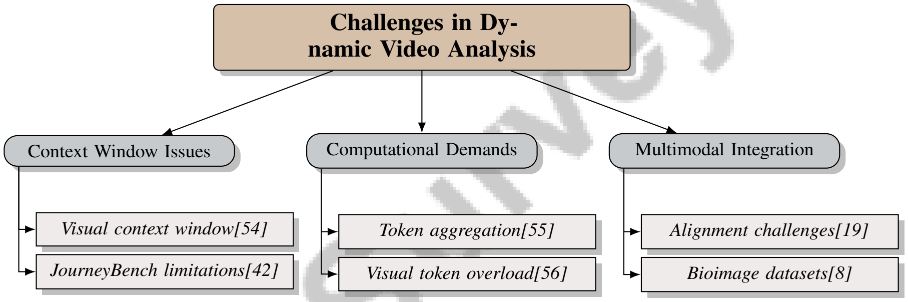


Figure 4: This figure illustrates the primary challenges in analyzing dynamic video content, focusing on context window issues, computational demands, and multimodal integration. The context window problems highlight difficulties in handling visual context and benchmark limitations. Computational challenges address token aggregation and visual token overload. Multimodal integration emphasizes alignment challenges and the need for bioimage datasets.

5 Evaluation Benchmarks

Establishing robust evaluation benchmarks is essential for assessing the performance of Multimodal Large Language Models (MLLMs). These benchmarks provide a structured framework to evaluate model capabilities across various tasks and scenarios, offering insights into their strengths and limitations. The following subsections explore the specifics of these benchmarks, illustrating their significance in advancing video understanding and analysis.

5.1 Overview of Evaluation Benchmarks

Evaluation benchmarks are crucial for assessing MLLMs in video understanding, providing structured frameworks for comprehensive evaluation across diverse tasks. InfiniBench, for instance, evaluates 1,219 videos and 108.2K question-answer pairs, highlighting the importance of robust benchmarks [34]. Similarly, the VALUE benchmark assesses VidL systems through tasks like text-to-video retrieval and video captioning, ensuring thorough evaluation of MLLMs [58]. ApolloBench streamlines evaluation with multiple-choice questions, reducing time while maintaining rigor [4].

In video quality assessment, benchmarks such as those for LMM-VQA, including MSVD and TVQA, are vital for evaluating MLLMs' video quality comprehension [7]. Evaluations on datasets like

Benchmark	Size	Domain	Task Format	Metric
InfiniBench[34]	1,219	Video Understanding	Question Answering	Accuracy, GPT-4 Score
VALUE[58]	262,000	Video-and-Language Understanding	Multi-Task Evaluation	AveR, CIDEr-D
ApolloBench[4]	400	Video Understanding	Multiple-Choice Question Answering	Accuracy, F1-score
PT[59]	11,620	Multimodal Perception	Video Question Answering	Avg. IoU, top-1 accuracy
MVBench[60]	200	Video Understanding	Question Answering	Accuracy, F1-score
MLLM-Bench[15]	420	Multimodal Evaluation	Open-ended Queries	Win Rate
FGVEdit[61]	11,112	Visual Question Answering	Knowledge Editing	Specificity, Reliability
Auto-Bench[62]	3,504,000	Vision-Language Processing	Question Answering	Accuracy, Agreement Rate

Table 2: Table presents a comprehensive comparison of various evaluation benchmarks used for assessing Multimodal Large Language Models (MLLMs) in video understanding and related tasks. The table details essential characteristics of each benchmark, including their size, domain focus, task format, and performance metrics employed to evaluate model efficacy. This overview provides insights into the diverse methodologies and evaluation criteria that contribute to the robust assessment of MLLMs.

MSVD-QA and ScienceQA underscore the significance of diverse benchmarks in assessing MLLM performance [1]. Zhao et al.'s benchmark challenges models with long videos, addressing temporal reasoning complexities [63].

Performance metrics like F1 score and CIDEr measure model accuracy in event localization and video summarization, as seen in the TimeChat framework [47]. Patraucean et al.'s benchmark evaluates multimodal models in reasoning and perceptual tasks using real-world data, emphasizing comprehensive evaluation frameworks [59].

Refining and expanding evaluation benchmarks are vital for advancing MLLMs, addressing existing methodology limitations, and enhancing understanding of MLLMs' capabilities in complex tasks. By introducing paradigms like MLLM-Bench, which uses per-sample criteria and human-like judgment, researchers can ensure MLLMs are equipped for diverse creative tasks, fostering progress in Artificial General Intelligence (AGI) [15, 48]. Table 2 provides a detailed overview of representative benchmarks utilized in evaluating Multimodal Large Language Models (MLLMs), highlighting their significance in advancing the field of video understanding.

5.2 Spatial and Temporal Evaluation Criteria

Evaluating spatial and temporal understanding in MLLMs is key to determining their effectiveness in processing complex video content. Metrics like accuracy and F1-score provide objective measures of performance in comprehension and reasoning based on video inputs [4]. These metrics ensure a comprehensive assessment of model capabilities in both multiple-choice and open-ended questions [34].

Metrics are selected to offer a coarse-to-fine assessment of MLLMs' proficiency in managing extended sequences, maintaining coherence over time, and expressing uncertainty, essential for robust video understanding [64, 5]. Qualitative metrics like Avg. IoU and top-1 accuracy evaluate precision in tracking and answering video-based questions, reflecting real-world complexities [59].

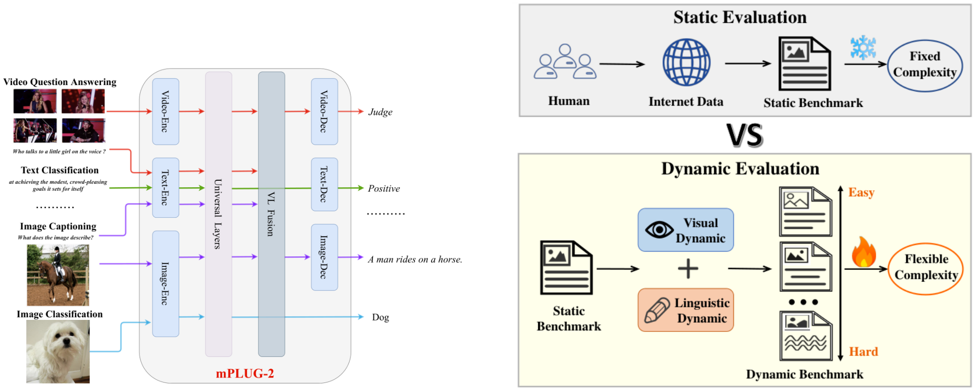
Evaluation criteria also include metrics for retrieval accuracy and captioning quality, such as AveR and CIDEr-D, crucial for tasks involving video retrieval and captioning [58]. Continuous refinement of evaluation criteria for spatial and temporal understanding advances MLLMs in video analysis, facilitating their application across various scenarios. Benchmarks like VidComposition and MLLM-as-a-Judge highlight MLLMs' limitations and identify improvement areas, driving progress in multimodal research and applications [21, 65, 45, 48].

5.3 Multimodal Evaluation Frameworks

Advanced evaluation frameworks are necessary to assess MLLMs' integration of diverse modalities, such as text, images, and audio, and their impact on model performance. These frameworks must surpass traditional methods, incorporating benchmarks that reflect capabilities in perception, reasoning, and domain-specific applications. By systematically addressing benchmark types, evaluation processes, and metrics, researchers can gain insights into model performance and drive MLLM advancements [48, 15, 26, 17, 49].

Muffin-Chihuahua evaluates MLLMs, including models like LLaVA and GPT-4V, highlighting the importance of evaluating multimodal adaptability [66]. Fu et al.'s hierarchical taxonomy categorizes benchmarks based on evaluation capabilities and applications, emphasizing the need for comprehensive assessment [48]. VideoVista evaluates models like Video-LLaMA, focusing on video understanding skills [67]. MMBench's CircularEval strategy and SEED-Bench's diverse questions ensure robust evaluation frameworks [68, 69].

NeedleVideo focuses on scalable evaluations of specific video skills, aiding model development for multimodal challenges [70]. VLB's dynamic evaluation samples capture the evolving nature of multimodal tasks [71]. By exploring multiple modalities, these frameworks guide MLLM development, ensuring effectiveness in diverse scenarios.



(a) mPLUG-2: A Unified Framework for Video, Text, and Image Classification and Answering[72]

(b) Dynamic Evaluation vs Static Evaluation in Image Retrieval Systems[71]

Figure 5: Examples of Multimodal Evaluation Frameworks

As shown in Figure 5, two prominent examples illustrate diverse methodologies in assessing complex systems integrating multiple data forms. "mPLUG-2" showcases an architecture integrating video, text, and image data through distinct encoding components, processed through Universal Layers for comprehensive analysis. Conversely, "Dynamic Evaluation vs Static Evaluation in Image Retrieval Systems" compares static and dynamic evaluation paradigms, emphasizing adaptable systems for modern multimodal data environments. These examples underscore the importance of sophisticated frameworks for effective evaluation and enhancement of multimodal systems [72, 71].

5.4 Future Directions in Benchmarking

The future of evaluation benchmarks for MLLMs involves significant advancements, emphasizing comprehensive and realistic assessments. Improving long context modeling will enhance benchmarks' ability to simulate real-world scenarios, capturing the multifaceted nature of multimodal reasoning [14]. Scaling benchmarks like MILEBENCH to accommodate larger contexts and additional modalities will develop robust MLLMs for complex tasks [44]. Expanding benchmarks to include diverse tasks and improving complex multimodal interaction evaluation will capture MLLMs' evolving capabilities [73].

Future research should focus on developing efficient multimodal models and exploring new applications across domains. Addressing current model limitations will enhance usability and ensure benchmarks remain relevant [27]. Refining benchmarks, expanding datasets, and exploring additional metrics are necessary to capture evolving MLLM capabilities [74].

Expanding datasets and refining metrics will apply benchmarks across different tasks, ensuring comprehensive assessment in diverse scenarios [63]. Exploring diverse tasks and scenarios will suggest potential benchmark improvements, enhancing effectiveness and applicability [59].

AI-generated, for reference only.

6 Applications and Case Studies

The transformative impact of Multimodal Large Language Models (MLLMs) is evident across various domains, demonstrating their versatility and efficacy in real-world applications. This section explores specific instances of MLLM applications, highlighting their contributions to fields such as healthcare, education, and beyond.

6.1 Applications of MLLMs in Various Domains

Multimodal Large Language Models (MLLMs) integrate diverse data types, including text, images, audio, and video, enabling them to excel in complex multimodal tasks like natural language understanding, visual reasoning, and recommendation systems. These models mark a significant advancement over traditional single-modality models, enhancing user experiences by adapting to evolving preferences and addressing current challenges while suggesting future development directions [32, 19, 17, 75]. MLLMs have notably improved dialogue systems and video generation tasks, as evidenced by the InternVid dataset, which evaluates multimodal reasoning capabilities.

In personalized applications, MLLMs are utilized in recommendation systems and image generation, offering tailored assistance across domains like healthcare and fashion [76]. This adaptability enhances user satisfaction by addressing specific needs and preferences.

Advancements in MLLM efficiency have expanded their viability for resource-sensitive tasks, broadening their practical applications [77]. Comprehensive benchmarks, such as those for edited videos, significantly contribute to video understanding, enabling better interpretation of complex content [78]. Datasets like MLVU provide a framework for long video understanding and enhance MLLM capabilities in diverse video analysis tasks [79].

MiniGPT4-Video exemplifies MLLMs' potential in enhancing video comprehension across domains, showcasing their ability to improve analysis capabilities [7]. These advancements underscore MLLMs' transformative potential in various applications, paving the way for more sophisticated AI systems in video understanding.

6.2 Medical and Healthcare Applications

MLLMs significantly advance medical and healthcare video analysis, enhancing diagnostic and treatment processes by integrating diverse data types such as medical imaging, patient records, and audio inputs. This integration supports clinical decision-making and patient engagement [80, 81, 18, 20, 82]. By fusing visual and textual information, MLLMs facilitate accurate diagnoses and personalized treatment plans.

In clinical settings, MLLMs automate medical image interpretation, combining visual data with patient histories to enhance diagnostic precision and reduce human error [38, 22, 83, 20]. They also support decision-making by synthesizing data from multiple sources.

MLLMs analyze surgical videos, aiding in identifying critical events and procedural steps. By integrating video data with textual descriptions, they enhance medical training and assessment, promoting adherence to best practices and improving patient outcomes [28, 25, 84, 57, 20].

In telemedicine, MLLMs facilitate remote consultations by synthesizing diverse data types to improve patient assessments, presenting opportunities for effective patient engagement and streamlined medical imaging analysis, despite challenges regarding data limitations and ethical considerations [21, 20, 23]. This capability is particularly beneficial in regions with limited healthcare access, enabling timely patient monitoring.

The deployment of MLLMs in healthcare underscores their potential to revolutionize the field, improving diagnostic accuracy, clinical training, and access to quality services. These advancements highlight MLLMs' essential role in fostering innovation and enhancing patient care by integrating diverse data types to support clinical decision-making and research outcomes while addressing implementation challenges and ethical concerns [81, 20, 82].

6.3 Educational and Instructional Video Analysis

MLLMs demonstrate significant potential in analyzing educational and instructional videos by integrating textual and visual data, thereby enhancing comprehension and overall learning experiences. Studies indicate that MLLMs can follow visual instructions, adapt to varying modalities, and leverage in-context learning, outperforming traditional single-modality systems in complex multimodal tasks. Their performance in visualization literacy has even surpassed human capabilities in certain analytical tasks, highlighting their robust application in education [85, 24, 26, 86, 17]. By facilitating information extraction and synthesis from video content, MLLMs improve the accessibility and effectiveness of educational materials.

In education, MLLMs generate comprehensive summaries and annotations of instructional videos, enhancing learners' understanding of complex concepts. Their advanced video analysis capabilities provide tailored insights that foster deeper comprehension. Benchmarks like VidComposition and Video-MME demonstrate MLLMs' potential in analyzing video compositions and integrating multimodal inputs, thereby improving engagement and information retention [21, 14, 24, 26]. By analyzing visual elements alongside audio or text, MLLMs create comprehensive notes that highlight key learning points, supporting diverse learning styles.

Moreover, MLLMs enhance educational video interactivity by enabling question-answering systems that allow learners to engage dynamically with content. These systems utilize large language models (LLMs) to interpret user queries and analyze relevant video segments, promoting active engagement and deeper material understanding. By employing learnable retrieval mechanisms to select pertinent video chunks, these systems effectively reduce computational noise, enhancing response accuracy [87, 10, 88, 30, 31].

Instructors also benefit from MLLMs in instructional video analysis, gaining insights into viewer engagement and comprehension. By examining learner interaction patterns, educators can refine their strategies and materials based on advanced multimodal analysis tools, enabling tailored approaches to meet diverse audience needs [21, 89, 90, 30].

MLLMs facilitate the development of multilingual instructional content, enhancing resource accessibility and inclusivity by enabling nuanced understanding of complex video compositions and integrating diverse materials. This capability enriches the educational experience, accommodating various learning styles [21, 25]. By processing and translating video content into multiple languages, MLLMs bridge language barriers, making educational resources more widely accessible.

The application of MLLMs in educational and instructional video analysis illustrates their potential to transform learning environments by enhancing content accessibility, interactivity, and personalization. These advancements underscore MLLMs' pivotal role in fostering innovation within educational technology, improving task performance through the integration of diverse data types, and ultimately leading to better learning outcomes across various educational settings [48, 24, 91, 26, 20].

6.4 Narrative and Long-Form Video Understanding

MLLMs have advanced significantly in understanding narrative and long-form videos, adeptly processing complex narratives through the integration of textual and visual information. Understanding such content requires discerning intricate temporal sequences and contextual subtleties, skills that MLLMs are progressively refining. Recent advancements, including the VidComposition benchmark, emphasize the need for MLLMs to analyze video compositions more intricately, revealing gaps in their capabilities compared to human understanding. Innovative approaches, such as the browse-and-concentrate paradigm and learnable retrieval-based models, are being developed to enhance MLLMs' comprehension of multimodal content, addressing context fusion challenges and efficient long video processing [21, 22, 25, 88, 26].

Recent developments, such as the TimeChat model, demonstrate MLLMs' ability to incorporate visual and timestamp information, improving long video understanding by accurately localizing and interpreting events over extended durations [47]. This capability is crucial for tasks requiring detailed narrative flow comprehension.

Models like MovieSeq utilize interleaved multimodal sequences to enhance narrative structure understanding within videos [10]. By interleaving visual and textual data, these models capture storylines and thematic elements more effectively, providing richer narrative analyses.

Challenges in processing long-form videos, such as maintaining context over extended durations and synthesizing information from multiple sources, are addressed through innovative architectures and training techniques in MLLMs. For instance, integrating visual tokens with subtitles in MiniGPT4-Video exemplifies how MLLMs enhance video comprehension through effective multimodal integration [7].

The development of comprehensive benchmarks like InfiniBench is crucial for evaluating MLLMs' capabilities in handling long video content, providing structured frameworks for assessing performance in narrative understanding tasks [34]. These benchmarks are essential for advancing research, ensuring MLLMs meet the demands of complex narrative analysis.

The application of MLLMs in narrative and long-form video understanding highlights their transformative potential in enhancing video comprehension. By employing advanced integration strategies and innovative model architectures, MLLMs are positioned to improve narrative content analysis in video, as evidenced by specialized benchmarks like VidComposition and StreamingBench, which assess MLLMs' abilities in understanding complex video compositions and real-time streaming scenarios, respectively. This progress aims to refine video comprehension and lays the groundwork for developing more sophisticated and reliable applications in video understanding [21, 24, 14, 53].

7 Challenges and Future Directions

Addressing the challenges of Multimodal Large Language Models (MLLMs) in video analysis requires understanding the complexities of video data integration. This involves examining the integration of diverse data modalities and maintaining coherent context over extended sequences. The following subsections discuss specific challenges in processing complex video data, highlighting computational inefficiencies and limitations of current methodologies that need to be addressed to enhance MLLM performance.

7.1 Handling Complex Video Data

MLLMs face significant challenges in processing complex video data due to the intricacies of integrating multiple data modalities while maintaining coherent context over long sequences. A major issue is computational inefficiency arising from the complexity of integrating modalities, often relying on discrete token representations that hinder scalability and real-time application in scenarios requiring simultaneous processing of text, image, and video data [4]. Current models, often dependent on pretrained encoders, focus on short-term patterns, lacking the ability to capture long-term dependencies and explicit temporal annotations essential for aligning spatial, temporal, and semantic information [50]. The reliance on human quality scores further restricts benchmark applicability to unlabeled datasets, complicating MLLM evaluation [35].

JourneyBench highlights the inadequacies of multimodal models in reasoning about unusual and fictional visual scenarios, while conflicts between image-level and region-level vision-language tasks challenge effective training and integration of visual information [2]. Efficiency gains from frameworks like AIM are undermined by lengthy textual labels, necessitating more efficient token management strategies [36]. Frameworks like MovieSeq may not accommodate all narrative complexities in visual input processing [10], highlighting the need for adaptable processing architectures. The scarcity of high-quality, annotated datasets remains a significant barrier, limiting model fine-tuning for specific tasks without overfitting [8]. Future research should focus on developing robust evaluation frameworks and innovative methodologies to enhance MLLM capabilities in handling complex video data, exploring longer contexts to capture human reasoning intricacies and the complementary nature of diverse datasets.

7.2 Integration with Computational Linguistics

Integrating MLLMs with computational linguistics is essential for enhancing analytical capabilities in complex multimodal tasks. This integration leverages language models' strengths to interpret and generate textual data, improving MLLM performance in video analysis. Verma et al. emphasize the importance of fine-tuning linguistic components to effectively process and integrate visual information [92]. The benchmark introduced by Zhang et al. identifies limitations in current multimodal models, emphasizing improved consistency across modalities [93]. Challenges such as reliance on translated datasets may introduce inaccuracies or cultural misinterpretations, underscoring the need for robust datasets and methodologies that accommodate linguistic diversity [94].

The assumption that LLMs utilize specialized attention heads for improved visual task performance, as proposed by Bi et al., represents a shift from traditional language-focused approaches [95]. This innovation highlights the potential for integrating computational linguistics with visual processing for nuanced video analysis. Tao et al. stress the importance of understanding MLLMs' representation capabilities to enhance performance in complex multimodal tasks [96]. Despite advancements, high computational requirements and challenges in achieving optimal model performance across diverse modalities remain, necessitating ongoing research to enhance MLLM and computational linguistics integration [27].

7.3 Scalability and Efficiency

Scalability and efficiency are critical for MLLMs' applicability across diverse domains. Models like mPLUG-2 excel in integrating multimodal data, but their complexity poses challenges for scalability and implementation in specific applications [72]. This complexity often leads to increased computational demands, hindering deployment in resource-constrained environments. Designing MLLMs requires careful consideration of trade-offs between model complexity and efficiency, significantly influencing performance across applications such as natural language processing and visual understanding. Balancing these trade-offs is vital for optimizing capabilities and practical deployment, especially in resource-constrained environments like edge computing [77, 15, 17, 48]. Strategies to enhance scalability include optimizing model architectures to reduce computational overhead while maintaining performance through techniques like model compression, efficient token management, and distributed computing.

Advanced training techniques, such as stage-wise instruction tuning and modular design, are crucial for improving scalability. These methods enhance instruction-following capabilities by reducing visual redundancy and optimizing data mixtures during training. Techniques like Visual-Modality Token Compression and dynamic input scaling help maintain high performance across diverse tasks, leading to better adaptability and efficiency in real-world applications [97, 41, 15, 43, 98]. The challenge of balancing efficiency with performance is further complicated by the need to process long-form video content, necessitating models to maintain coherence over extended sequences. Innovative methodologies that efficiently manage and integrate diverse data modalities are essential for analyzing complex video data. MLLMs have shown potential in understanding complex video compositions through benchmarks like VidComposition, which evaluates MLLMs' capabilities in interpreting nuanced interactions among visual elements, revealing performance gaps between human understanding and model capabilities. Resource-efficient strategies, such as the RED-VILLM pipeline, facilitate the transition from Image LLMs to Video LLMs by incorporating temporal information, enhancing MLLM effectiveness in video analysis while minimizing resource requirements [21, 23].

Continuous refinement of scalability and efficiency strategies is essential for advancing MLLM capabilities. By optimizing model architectures and refining training techniques, researchers can improve the deployment and versatility of MLLMs, ensuring effectiveness in various applications, particularly in resource-sensitive environments. Ongoing evaluations and the development of innovative benchmarks will further guide these improvements, fostering MLLM advancement in tackling complex, real-world tasks [16, 26, 15, 48].

7.4 Innovative Research Directions

Advancing MLLMs in video analysis requires exploring innovative research directions that address existing challenges while pushing the boundaries of current capabilities. One promising avenue is refining soft prompting methods, as exemplified by the LION framework, which could be enhanced by investigating additional multimodal tasks to broaden its application scope [2]. This approach underscores the potential for improving model adaptability and performance across diverse video analysis scenarios. Enhancing model architecture to increase semantic density is another critical area for future research. The TimeChat model suggests a need for more diverse instruction-tuning data to expand time-related applications [47]. This focus on semantic richness and diversity in training data is crucial for improving MLLM temporal reasoning capabilities, enabling better handling of complex video content.

Joint training of large language models with encoders and developing richer datasets are essential for fostering better multimodal alignment and temporal reasoning [50]. Expanding benchmarks like ApolloBench to encompass a wider range of video understanding tasks and refining evaluation metrics will ensure comprehensive assessments of model performance [4]. Optimizing demonstration caching to reduce storage requirements and enhance efficiency is another promising research direction [36]. Expanding benchmarks such as JourneyBench to include more diverse visual scenarios and refining evaluation metrics will significantly advance the field [42].

Furthermore, integrating multilingual VidL datasets and diagnostic datasets for deeper analysis will expand MLLM applicability across various languages and cultural contexts [58]. Incorporating knowledge distillation techniques and improving robustness in fact-checking capabilities are vital for future MLLM development [5]. Pursuing these innovative research directions will achieve significant advancements, ensuring MLLMs remain at the forefront of video understanding and analysis, capable of addressing increasingly complex multimodal challenges.

8 Conclusion

The exploration of Multimodal Large Language Models (MLLMs) unveils their significant potential in advancing video understanding by seamlessly integrating multiple data modalities. Through frameworks like MR-MLLM, these models have set new standards in multimodal comprehension and vision tasks, showcasing their ability to process complex video content, as demonstrated by MiniGPT4-Video. Such advancements highlight the pivotal role of MLLMs in enhancing video analysis capabilities.

Moreover, the survey underscores the necessity of robust evaluation methods to guide the evolution of MLLMs. The CODIS framework, for instance, identifies existing limitations in context-dependent visual comprehension, suggesting pathways for future model improvements. The applicability of MLLMs extends beyond video analysis, with promising implications in fields like bioimage analysis, where they offer intelligent solutions adaptable to diverse research contexts.

Nonetheless, challenges remain, particularly in capturing intricate human-object interactions and temporal dynamics within video data. Current benchmarks, such as EgoPlanBench2, emphasize the need for enhanced ensemble techniques and adaptability to diverse video formats. Future research directions should focus on integrating additional modalities and refining the understanding of relationships between examples, thereby improving the design of supervised retrievers for multimodal in-context learning.

AI-generated, for reference only.

References

[1] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023.

[2] Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. Lion : Empowering multimodal large language model with dual-level visual knowledge, 2023.

[3] Tao Wu, Mengze Li, Jingyuan Chen, Wei Ji, Wang Lin, Jinyang Gao, Kun Kuang, Zhou Zhao, and Fei Wu. Semantic alignment for multimodal large language models, 2024.

[4] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, Serena Yeung-Levy, and Xide Xia. Apollo: An exploration of video understanding in large multimodal models, 2024.

[5] Jiahui Geng, Yova Kementchedjhieva, Preslav Nakov, and Iryna Gurevych. Multimodal large language models to support real-world fact-checking, 2024.

[6] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models, 2024.

[7] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens, 2024.

[8] Shanghang Zhang, Gaole Dai, Tiejun Huang, and Jianxu Chen. Multimodal large language models for bioimage analysis, 2024.

[9] Fuwen Luo, Chi Chen, Zihao Wan, Zhaolu Kang, Qidong Yan, Yingjie Li, Xiaolong Wang, Siyu Wang, Ziyue Wang, Xiaoyue Mi, Peng Li, Ning Ma, Maosong Sun, and Yang Liu. Codis: Benchmarking context-dependent visual comprehension for multimodal large language models, 2024.

[10] Kevin Qinghong Lin, Pengchuan Zhang, Difei Gao, Xide Xia, Joya Chen, Ziteng Gao, Jinheng Xie, Xuhong Xiao, and Mike Zheng Shou. Learning video context as interleaved multimodal sequences, 2024.

[11] Zhanyu Wang, Longyue Wang, Zhen Zhao, Minghao Wu, Chenyang Lyu, Huayang Li, Deng Cai, Luping Zhou, Shuming Shi, and Zhaopeng Tu. Gpt4video: A unified multimodal large language model for lnstruction-followed understanding and safety-aware generation, 2024.

[12] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability, 2023.

[13] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding, 2024.

[14] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

[15] Wentao Ge, Shunian Chen, Guiming Hardy Chen, Junying Chen, Zhihong Chen, Nuo Chen, Wenya Xie, Shuo Yan, Chenghao Zhu, Ziyue Lin, Song Dingjie, Xidong Wang, Anningzhe Gao, Zhang Zhiyi, Jianquan Li, Xiang Wan, and Benyou Wang. Mllm-bench: Evaluating multimodal llms with per-sample criteria, 2024.

[16] Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning, 2024.

AI-generated, for reference only.

[17] Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, Xin Zhang, Wei Zhang, Dinggang Shen, Tianming Liu, and Shu Zhang. A comprehensive review of multimodal large language models: Performance and challenges across different tasks, 2024.

[18] Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, Yong Liu, Jing Shao, Hui Xiong, and Xuming Hu. Explainable and interpretable multimodal large language models: A comprehensive survey, 2024.

[19] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models, 2024.

[20] Qian Niu, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Junyu Liu, Benji Peng, Tianyang Wang, Yunze Wang, Silin Chen, and Ming Liu. From text to multimodality: Exploring the evolution and impact of large language models in medical practice, 2024.

[21] Yunlong Tang, Junjia Guo, Hang Hua, Susan Liang, Mingqian Feng, Xinyang Li, Rui Mao, Chao Huang, Jing Bi, Zeliang Zhang, Pooyan Fazli, and Chenliang Xu. Vidcomposition: Can mllms analyze compositions in compiled videos?, 2024.

[22] Ziyue Wang, Chi Chen, Yiqi Zhu, Fuwen Luo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Maosong Sun, and Yang Liu. Browse and concentrate: Comprehending multimodal content via prior-llm context fusion, 2024.

[23] Suyuan Huang, Haoxin Zhang, Yan Gao, Yao Hu, and Zengchang Qin. From image to video, what do we need in multimodal llms?, 2024.

[24] Enis Berk Çoban, Michael I. Mandel, and Johanna Devaney. What do mllms hear? examining reasoning with text and sound components in multimodal large language models, 2024.

[25] Shukang Yin, Chaoyou Fu, Sirui Zhao, Yunhang Shen, Chunjiang Ge, Yan Yang, Zuwei Long, Yuhan Dai, Tong Xu, Xing Sun, Ran He, Caifeng Shan, and Enhong Chen. T2vid: Translating long text into multi-image is the catalyst for video-llms, 2024.

[26] Yang Luo, Zangwei Zheng, Zirui Zhu, and Yang You. How does the textual information affect the retrieval of multimodal in-context learning?, 2024.

[27] Soyeon Caren Han, Feiqi Cao, Josiah Poon, and Roberto Navigli. Multimodal large language models and tunings: Vision, language, sensors, audio, and beyond, 2024.

[28] Nafisa Hussain. Multimodal language models for domain-specific procedural video summarization, 2024.

[29] Yapeng Tian, Chenxiao Guan, Justin Goodman, Marc Moore, and Chenliang Xu. An attempt towards interpretable audio-visual video captioning, 2018.

[30] This cvpr 2020 paper is the open.

[31] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

[32] Tianyi Bai, Hao Liang, Binwang Wan, Yanran Xu, Xi Li, Shiyu Li, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, Ping Huang, Jiulong Shan, Conghui He, Binhang Yuan, and Wentao Zhang. A survey of multimodal large language model from a data-centric perspective, 2024.

[33] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE, 2023.

[34] Kirolos Ataallah, Chenhui Gou, Eslam Abdelrahman, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. Infinibench: A comprehensive benchmark for large multimodal models in very long video understanding, 2024.

[35] Tianhe Wu, Kede Ma, Jie Liang, Yujiu Yang, and Lei Zhang. A comprehensive study of multimodal large language models for image quality assessment, 2024.

[36] Jun Gao, Qian Qiao, Ziqiang Cao, Zili Wang, and Wenjie Li. Aim: Let any multi-modal large language models embrace efficient in-context learning, 2024.

[37] Taylor Arnold and Lauren Tilton. Explainable search and discovery of visual cultural heritage collections with multimodal large language models, 2024.

[38] Wei Chen, Lin Li, Yongqi Yang, Bin Wen, Fan Yang, Tingting Gao, Yu Wu, and Long Chen. Comm: A coherent interleaved image-text dataset for multimodal understanding and generation, 2024.

[39] Yinheng Li, Han Ding, and Hang Chen. Data processing techniques for modern multimodal models, 2024.

[40] Marco AF Pimentel, Clément Christophe, Tathagata Raha, Prateek Munjal, Praveen K Kanithi, and Shadab Khan. Beyond metrics: A critical analysis of the variability in large language model evaluation frameworks, 2024.

[41] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. Mm1: Methods, analysis insights from multimodal llm pre-training, 2024.

[42] Zhecan Wang, Junzhang Liu, Chia-Wei Tang, Hani Alomari, Anushka Sivakumar, Rui Sun, Wenhao Li, Md. Atabuzzaman, Hammad Ayyubi, Haoxuan You, Alvi Ishmam, Kai-Wei Chang, Shih-Fu Chang, and Chris Thomas. Journeybench: A challenging one-stop vision-language understanding benchmark of generated images, 2024.

[43] Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, Sam Dodge, Keen You, Zhen Yang, Aleksei Timofeev, Mingze Xu, Hong-You Chen, Jean-Philippe Fauconnier, Zhengfeng Lai, Haoxuan You, Zirui Wang, Afshin Dehghan, Peter Grasch, and Yinfei Yang. Mm1.5: Methods, analysis insights from multimodal llm fine-tuning, 2024.

[44] Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context, 2024.

[45] Wen Wen, Yilin Wang, Neil Birkbeck, and Balu Adsumilli. An ensemble approach to short-form video quality assessment using multimodal llm, 2024.

[46] 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.

[47] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024.

[48] Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, Caifeng Shan, and Ran He. Mme-survey: A comprehensive survey on evaluation of multimodal llms, 2024.

[49] Jiaxing Huang and Jingyi Zhang. A survey on evaluation of multimodal large language models, 2024.

[50] Xi Ding and Lei Wang. Do language models understand time?, 2025.

[51] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent, 2024.

[52] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

[53] Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding, 2024.

[54] Hongchen Wei and Zhenzhong Chen. Visual context window extension: A new perspective for long video understanding, 2024.

[55] Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding, 2024.

[56] Zhuqiang Lu, Zhenfei Yin, Mengwei He, Zhihui Wang, Zicheng Liu, Zhiyong Wang, and Kun Hu. B-vllm: A vision large language model with balanced spatio-temporal tokens, 2024.

[57] Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Jameel Hassan, Muzammal Naseer, Federico Tombari, Fahad Shahbaz Khan, and Salman Khan. How good is my video lmm? complex video reasoning and robustness evaluation suite for video-lmms, 2024.

[58] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, et al. Value: A multi-task benchmark for video-and-language understanding evaluation. *arXiv preprint arXiv:2106.04632*, 2021.

[59] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761, 2023.

[60] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2024.

[61] Zhen Zeng, Leijiang Gu, Xun Yang, Zhangling Duan, Zenglin Shi, and Meng Wang. Visual-oriented fine-grained knowledge editing for multimodal large language models, 2024.

[62] Yuanfeng Ji, Chongjian Ge, Weikai Kong, Enze Xie, Zhengying Liu, Zhengguo Li, and Ping Luo. Large language models as automated aligners for benchmarking vision-language models, 2023.

[63] Tiancheng Zhao, Qianqian Zhang, Kyusong Lee, Peng Liu, Lu Zhang, Chunxin Fang, Jiajia Liao, Kelei Jiang, Yibo Ma, and Ruochen Xu. Omchat: A recipe to train multimodal language models with strong long context and video understanding, 2024.

[64] Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models, 2025.

[65] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark, 2024.

[66] Yue Fan, Jing Gu, Kaiwen Zhou, Qianqi Yan, Shan Jiang, Ching-Chen Kuo, Xinze Guan, and Xin Eric Wang. Muffin or chihuahua? challenging multimodal large language models with multipanel vqa, 2024.

[67] Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. Videovista: A versatile benchmark for video understanding and reasoning, 2024.

[68] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[69] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023.

[70] Zijia Zhao, Haoyu Lu, Yuqi Huo, Yifan Du, Tongtian Yue, Longteng Guo, Bingning Wang, Weipeng Chen, and Jing Liu. Needle in a video haystack: A scalable synthetic evaluator for video mllms, 2024.

[71] Yue Yang, Shuibai Zhang, Wenqi Shao, Kaipeng Zhang, Yi Bin, Yu Wang, and Ping Luo. Dynamic multimodal evaluation with flexible complexity by vision-language bootstrapping, 2024.

[72] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. mplug-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning*, pages 38728–38748. PMLR, 2023.

[73] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024.

[74] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.

[75] Yuyang Ye, Zhi Zheng, Yishan Shen, Tianshu Wang, Hengruo Zhang, Peijun Zhu, Runlong Yu, Kai Zhang, and Hui Xiong. Harnessing multimodal large language models for multimodal sequential recommendation, 2025.

[76] Junda Wu, Hanjia Lyu, Yu Xia, Zhehao Zhang, Joe Barrow, Ishita Kumar, Mehrnoosh Mirtaheri, Hongjie Chen, Ryan A. Rossi, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, Jiuxiang Gu, Nesreen K. Ahmed, Yu Wang, Xiang Chen, Hanieh Deilamsalehy, Namyong Park, Sungchul Kim, Huanrui Yang, Subrata Mitra, Zhengmian Hu, Nedim Lipka, Dang Nguyen, Yue Zhao, Jiebo Luo, and Julian McAuley. Personalized multimodal large language models: A survey, 2024.

[77] Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Efficient multimodal large language models: A survey, 2024.

[78] Lu Xu, Sijie Zhu, Chunyuan Li, Chia-Wen Kuo, Fan Chen, Xinyao Wang, Guang Chen, Dawei Du, Ye Yuan, and Longyin Wen. Beyond raw videos: Understanding edited videos with large multimodal model, 2024.

[79] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

[80] Xiaoshuang Huang, Haifeng Huang, Lingdong Shen, Yehui Yang, Fangxin Shang, Junwei Liu, and Jia Liu. A refer-and-ground multimodal large language model for biomedicine, 2024.

[81] Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. A comprehensive survey of large language models and multimodal large language models in medicine, 2024.

[82] Xiaoshuang Huang, Lingdong Shen, Jia Liu, Fangxin Shang, Hongxiang Li, Haifeng Huang, and Yehui Yang. Towards a multimodal large language model with pixel-level insight for biomedicine, 2025.

[83] Zhelun Shi, Zhipin Wang, Hongxing Fan, Zaibin Zhang, Lijun Li, Yongting Zhang, Zhenfei Yin, Lu Sheng, Yu Qiao, and Jing Shao. Assessment of multimodal large language models in alignment with human values, 2024.

[84] Yuchen Yang and Yingxuan Duan. Towards holistic language-video representation: the language model-enhanced msr-video to text dataset, 2024.

[85] Xiujun Li, Yujie Lu, Zhe Gan, Jianfeng Gao, William Yang Wang, and Yejin Choi. Text as images: Can multimodal large language models follow printed instructions in pixels?, 2024.

[86] Zhimin Li, Haichao Miao, Valerio Pascucci, and Shusen Liu. Visualization literacy of multimodal large language models: A comparative study, 2024.

[87] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation, 2024.

[88] Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Long video understanding with learnable retrieval in video-language models, 2025.

[89] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2917–2927, 2022.

[90] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey, 2024.

[91] Yunxin Li, Baotian Hu, Wei Wang, Xiaochun Cao, and Min Zhang. Towards vision enhancing llms: Empowering multimodal knowledge storage and sharing in llms, 2023.

[92] Gaurav Verma, Minje Choi, Kartik Sharma, Jamelle Watson-Daniels, Sejoon Oh, and Srijan Kumar. Cross-modal projection in multimodal llms doesn't really project visual attributes to textual space, 2024.

[93] Xiang Zhang, Senyu Li, Ning Shi, Bradley Hauer, Zijun Wu, Grzegorz Kondrak, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. Cross-modal consistency in multimodal large language models, 2024.

[94] Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. Peacock: A family of arabic multimodal large language models and benchmarks, 2024.

[95] Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, and Chenliang Xu. Unveiling visual perception in language models: An attention head analysis approach, 2024.

[96] Mingxu Tao, Quzhe Huang, Kun Xu, Liwei Chen, Yansong Feng, and Dongyan Zhao. Probing multimodal large language models for global and local semantic representations, 2024.

[97] Yonghui Wang, Wengang Zhou, Hao Feng, and Houqiang Li. Adaptvision: Dynamic input scaling in mllms for versatile scene understanding, 2024.

[98] Te Yang, Jian Jia, Xiangyu Zhu, Weisong Zhao, Bo Wang, Yanhua Cheng, Yan Li, Shengyuan Liu, Quan Chen, Peng Jiang, Kun Gai, and Zhen Lei. Enhancing instruction-following capability of visual-language models by reducing image redundancy, 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.