A Vid-LLM architectures systemic analysis for the integration of multimodal video understanding

Svitlana Antoshchuk[†], Olena Arsirii[†], Mykola Hodovychenko[†], Anatolii Nikolenko[*,†] and Oksana Babilunha[†]

Odesa Polytechnic National University, Shevchenko Avenue 1, 65044 Odesa, Ukraine

Abstract

Recent advances in Large Language Models (LLMs) enable Video-Language Models (Vid-LLMs) for complex spatiotemporal video understanding. A systemic analysis of modern Vid-LLM architectures is presented, highlighting three main categories based on input processing strategies: Analyzer + LLM (relying on symbolic outputs), Embedder + LLM (using visual representations), and Hybrid frameworks as a combination of the first two. We analyzed their design principles, functional roles, and applications (captioning, QA, localization, agents). Challenges in long-context modeling, video tokenization, grounded reasoning, and integration with external tools are discussed. In conclusion, future research directions for improving Vid-LLM scalability, interpretability, and robustness are substantiated.

Keywords

Vid-LLMs, Video Comprehension, Multimodal Architectures, Spatio-temporal Reasoning, Video Analysis.

1. Introduction

The swift expansion of video content across digital platforms, such as social media, entertainment, surveillance, and autonomous systems, has generated an increased demand for intelligent systems capable of autonomously analyzing and comprehending complex visual data. Video comprehension, encompassing the identification of objects, actions, events, and the inference of high-level semantics over time, represents a core challenge in the fields of computer vision and artificial intelligence [1]. Conventional approaches, such as manual feature engineering and early neural networks, laid the groundwork for significant progress; however, they have proven inadequate in fully conveying the complexity and diversity inherent in real-world video footage [2].

In the last ten years, deep learning has made models much better at handling spatio-temporal data. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and more recently, Transformer-based architectures, have been successful in recognizing actions, classifying videos, adding captions, and finding the right time. Self-supervised learning has sped up this advancement even further by making it possible to train strong video encoders without a lot of human annotation. But these models are frequently just good at certain tasks and don't have the generalization and reasoning skills needed to interpret videos that are more abstract and include more than one phase [3].

Concurrently, Large Language Models (LLMs) such as GPT-4 [4], PaLM [5], and LLaMA [6] have attained pioneering results in natural language processing tasks. These models demonstrate emergent capabilities, such as few-shot learning, instruction adherence, and advanced reasoning, by utilizing extensive text corpora during the pretraining phase. Recent initiatives have commenced to investigate the integration of large language models with video data, leading to the

[*] *AIT&AIS'2025: International Scientific Workshop on Applied Information Technologies and Artificial Intelligence Systems, December 18–19 2025, Chernivtsi, Ukraine*

[1,*] Corresponding author.

[†] These authors contributed equally.

asg@op.edu.ua (S. Antoshchuk); e.arsiriy@gmail.com (O. Arsirii); hodovychenko@op.edu.ua (M. Hodovychenko); anatolyn@ukr.net (A. Nikolenko); babilunga.onpu@gmail.com (O. Babilunha)

0000-0002-9346-145X (S. Antoshchuk); 0000-0001-8130-9613 (O. Arsirii); 0000-0001-5422-3048 (M. Hodovychenko); 0000-0002-9849-1797 (A. Nikolenko); 0000-0001-6431-3557 (O. Babilunha)

emergence of a novel category of video-language models (Vid-LLMs) [7]. These systems integrate vision encoders and language models to enable multimodal video comprehension, capable of answering questions about videos, generating descriptions, identifying temporal events, and performing commonsense reasoning based on visual input.

Vid-LLMs offer a unified interface for executing various video comprehension tasks through prompting or in-context learning, generally without requiring extensive retraining. Their versatility and applicability across various domains render them valuable for purposes including robotics, surveillance, education, and content moderation.

The objective of this study is to perform a systematic analysis and classification of Video- Language Model (Vid-LLM) architectures to identify and assess the principal integration strategies between the visual and language modalities that characterize the capacity of Vid-LLMs for intricate multimodal video comprehension. The following goals are addressed:

1. Systematize the most advanced methods for developing Vid-LLMs by categorizing them according to architectural paradigms (Analyzer + LLM, Embedder + LLM, and Hybrid), and evaluate their fundamental design principles.
2. Assess the capabilities of Vid-LLMs across a broad spectrum of applications, including query answering, temporal localization, and agentic reasoning.
3. Identify the technical limitations and assessment criteria associated with video tokenization, long-context modeling, and interpretability.
4. Develop a strategic roadmap and delineate future research directions aimed at creating more robust, scalable, and efficient systems for visual environment integration.

2. Trends and milestones in video comprehension methods

The field of video comprehension has undergone substantial development in the last two decades, driven by advancements in computer vision, machine learning, and, more recently, multimodal artificial intelligence. The increasing volume of video data across diverse domains, including entertainment, social media, surveillance, and autonomous systems, has heightened the necessity for efficient and scalable methods for its interpretation and analysis. This paper analyzes the historical development of video comprehension techniques.

2.1. Early techniques

The initial phase of video comprehension involved manually crafted feature extraction techniques and conventional machine learning algorithms. Spatial characteristics have traditionally been obtained using descriptors such as Scale-Invariant Feature Transform (SIFT) [8], Speeded-Up Robust Features (SURF) [9], and Histogram of Oriented Gradients (HOG) [10], which aid in identifying and representing key visual patterns within discrete frames. Techniques including optical flow, background removal, and Improved Dense Trajectories (IDT) [11] were employed to characterize motion and temporal dynamics.

Temporal dependencies in video sequences have frequently been examined through statistical models, notably Hidden Markov Models (HMMs) [12], which enabled the recognition of sequential patterns. Conventional machine learning models, such as Support Vector Machines (SVMs) [13], Decision Trees [14], and Random Forests [15], have been extensively utilized for classification and recognition tasks. Furthermore, unsupervised methods including cluster analysis and dimensionality reduction techniques such as Principal Component Analysis (PCA) were employed to categorize video segments and decrease computational complexity.

The techniques provided valuable insights into video analysis; however, their applicability in other contexts was limited, and they faced challenges in scaling, particularly when dealing with complex, high-dimensional, or extended-duration videos. This prompted the exploration of more dependable methods, ultimately resulting in the adoption of deep learning-based techniques.

2.2. First-generation neural video models

Initial neural video models represented a substantial transition from conventional handmade methods by using deep learning architectures, especially convolutional and recurrent neural networks. Early models like DeepVideo [16] used 3D Convolutional Neural Networks (CNNs) [17] to derive visual features from video frames; however, they failed to surpass handmade features owing to insufficient motion representation. To overcome this problem, two-stream networks were developed, integrating RGB frame data with motion information (e.g., optical flow) to more effectively capture temporal dynamics.

Recurrent Neural Networks (RNNs) [18], particularly Long Short-Term Memory (LSTM) [19] networks, were used to analyze sequential data and improve the representation of long-range temporal relationships. Temporal Segment Networks (TSN) [20] consolidated data from poorly sampled segments to facilitate efficient analysis of long-form videos. Additional advances, including Fisher Vectors [21] and Bi-linear pooling [22], were used to enhance video-level representations.

The advent of 3D CNNs, including C3D and Inflated 3D ConvNets (I3D) [23], facilitated the integrated modeling of spatial and temporal data via volumetric convolutions.

The models demonstrated impressive performance on benchmarks such as UCF-101 [24] and HMDB51 [25], resulting in the adaptation of well-known 2D architectures (e.g., ResNet, SENet) into 3D formats (e.g., R3D, MFNet, STC) [26]. To enhance computational efficiency, decomposed convolution methods (e.g., S3D, ECO, P3D) [27] divide 3D operations into separable 2D and 1D convolutions.

Subsequent progress involved the development of long-range temporal modeling techniques (e.g., LTC, T3D, Non-local Networks, V4D) [28] and the introduction of efficient architectures such as SlowFast and X3D [29]. The incorporation of Vision Transformers (ViT) [30] has spurred the development of models including TimeSformer [31], ViViT [32], and MViT [33]. These models substitute convolutional operations with attention mechanisms, thereby providing enhanced scalability and improved temporal reasoning abilities for intricate video understanding applications.

2.3. Unsupervised pretraining for video understanding

Self-supervised pretraining for movies is a big step forward in video understanding because it lets models learn complete and generalizable representations from huge amounts of unprocessed video data. This method makes it easier to switch between jobs and lessens the need for notes that are specific to each task. VideoBERT [34] was a groundbreaking and well-known model that used hierarchical k-means clustering to tokenize video features and masked modeling to get representations that could go both ways. This model could be improved so that it does better at things like recognizing actions and adding captions to videos.

Subsequently, various approaches employed the pretraining-finetuning paradigm, incorporating innovations in architecture and training objectives. Models such as ActBERT, Spatio-temporalMAE, OmniMAE, VideoMAE, and MotionMAE have investigated masked video modeling and multimodal learning [35]. Others, including MaskFeat and CLIP-ViP, concentrated on contrastive learning and vision-language alignment [36]. These models incorporated mechanisms for reconstructing or predicting obscured video segments, aligning visual and textual modalities, or generating latent feature representations that encode both temporal and semantic information.

Self-supervised models have markedly enhanced performance on standard video benchmarks and exhibited robust generalization to tasks such as video classification, summarization, captioning, and question answering [37]. They also endorsed cross-modal learning, whereby coupled video and language data enabled the development of video-language models proficient in multimodal reasoning. This phase established the foundation for the integration of pretrained visual models with large language models in subsequent systems.

2.4. LLM-based approaches to video comprehension

The incorporation of Large Language Models (LLMs) into video comprehension represents the most recent and significant advancement in the domain (Fig. 1). Large Language Models, like ChatGPT and GPT-4, pretrained on extensive text corpora, exhibit robust in-context learning, instruction adherence, and reasoning ability. Their application to video comprehension represents a paradigm change by framing intricate video interpretation difficulties as language modeling challenges, often without necessitating considerable task-specific fine-tuning.

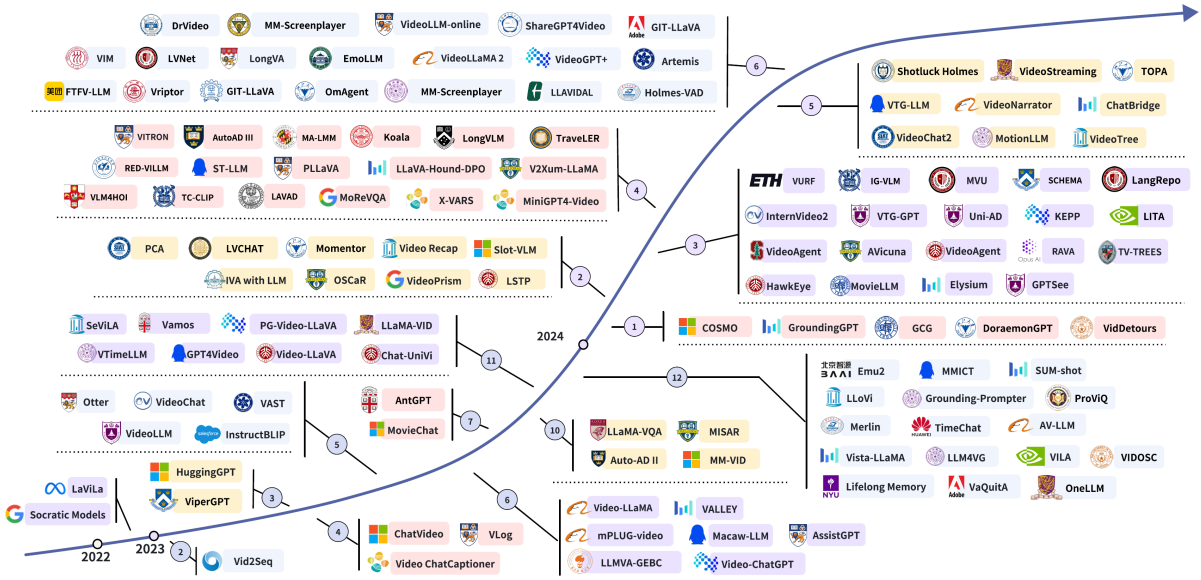

Figure 1: A visual timeline charting the development of video comprehension technologies powered by large language models. (Source: [3]).

Large language models (LLMs) are capable of processing textual representations obtained from video content or engaging with visual information via multimodal encoders. Utilizing these capabilities, systems like Visual-ChatGPT and other Vid-LLMs have been created to execute open-ended video reasoning, generate captions, respond to video-related inquiries, and invoke external vision APIs or tools based on prompts. This facilitates a dynamic and flexible comprehension of visual scenes through natural language. Instruction tuning and prompt engineering are essential for adapting large language models to perform various video-related tasks. These models demonstrate emergent capabilities, enabling them to perform multi-granularity reasoning – abstract, temporal, and spatio-temporal – through the integration of visual and commonsense knowledge. In contrast to earlier models designed for specific tasks, LLM-based video understanding systems exhibit the ability to generalize across various tasks through unified interfaces and few-shot or zero-shot learning [38].

Combining LLMs with video analysis opens the door to more scalable and human-like video comprehension systems that can handle multimodal problems in the real world in fields like robotics, education, entertainment, and surveillance. This development marks a move toward instruction-driven, general-purpose video intelligence.

3. Problem statement

Let a video be defined as a sequence of visual frames over time:

$$V=\{f_1,f_2,\ldots,f_T\},\quad f_t\in R^{H\times V\times C},\tag{1}$$

where each frame f_t is an RGB image of spatial resolution $H \times V$ with $C = 3$ channels, and T denotes the temporal length of the video.

Let an optional multimodal query Q be a sequence of language tokens

$$Q=\{q_1,q_2,\ldots,q_L\},\quad q_i\in V,\tag{2}$$

where V is the vocabulary space of the language model, and L is the length of the query or prompt.

The goal of video comprehension with Large Language Models (LLMs) is to define a function:

$$F_{\Theta}:(V,Q)\rightarrow A, \tag{3}$$

where F_{θ} is a parameterized model (e.g., a Vid-LLM) that maps the video and optional query to a structured output A, such as:

1. A natural language sequence: $A \in V^{*}$.
2. A classification label: $A \in Y$.
3. Or continuous-valued predictions (e.g., timestamps, coordinates): $A \in R^{d}$.

To learn this mapping, the system typically consists of:

1. Video encoder φ:

$$\phi : V \rightarrow v=\{v_1, v_2, \ldots, v_T\} \quad v_t \in R^{d_v}. \tag{4}$$

2. Language encoder (optional) ψ:

$$\psi : Q \rightarrow q=\{q_1, q_2, \ldots, q_L\} \quad q_i \in R^{d_q}. \tag{5}$$

3. Multimodal fusion function that aligns and integrates video and language features into a unified space interpretable by the LLM.
4. LLM core V : an autoregressive transformer that models the conditional probability distribution:

$$M(x_{1:i-1})=p(x_i \vee x_{1:i-1}), \tag{6}$$

where $x_i \in V \cup \{specialtokens\}$ and $x_{1:i-1}$ may include fused visual-linguistic context.

The training objective is to minimize a task-specific loss L, for instance:

$$\theta^{*}=argmin(E_{(V,Q,A^{*})\sim D}[L(F_{\theta}(V,Q)F^{*})]), \tag{7}$$

where A^{*} is the ground truth target from dataset D , and θ includes parameters from the encoders, fusion module, and (optionally) the LLM.

This formulation encapsulates various subtasks, including:

1. Captioning: $A \in V^{*}$ or Y.
2. Localization: $A = (t_s, t_e)$, t_s, t_e, $\in [1, T]$.
3. Tracking: $A=\{|(x_t, y_t)|\}_{t=1}^{T}$, etc.

The primary challenge lies in bridging the modality gap between continuous spatio-temporal visual signals and discrete symbolic reasoning in LLMs, while maintaining scalability, generalization, and data efficiency.

4. Vid-LLMs classification

Based on the strategy used to process input video data, we divide Vid-LLMs into three main categories.

4.1. Video Analyzer + LLM

The Video Analyzer + LLM architecture exemplifies a modular strategy for video comprehension, where the video content is initially handled by a specialized video analyzer module that extracts interpretable intermediate representations in textual format (Fig. 2). The outputs generally encompass video captions, detailed temporal captions, object tracking data, audio transcriptions

(through ASR), or subtitle text (through OCR) [39]. The resulting textual descriptions are subsequently provided as input to a Large Language Model (LLM), which conducts high-level reasoning, question answering, or task-specific inference based on the structured input [40].

This design effectively reformulates video comprehension as a text-based reasoning task, allowing LLMs to operate without the need for direct visual or spatio-temporal input processing. As a result, it leverages the zero-shot and in-context learning capabilities of pretrained LLMs while avoiding the computational overhead and training complexity of end-to-end multimodal models.

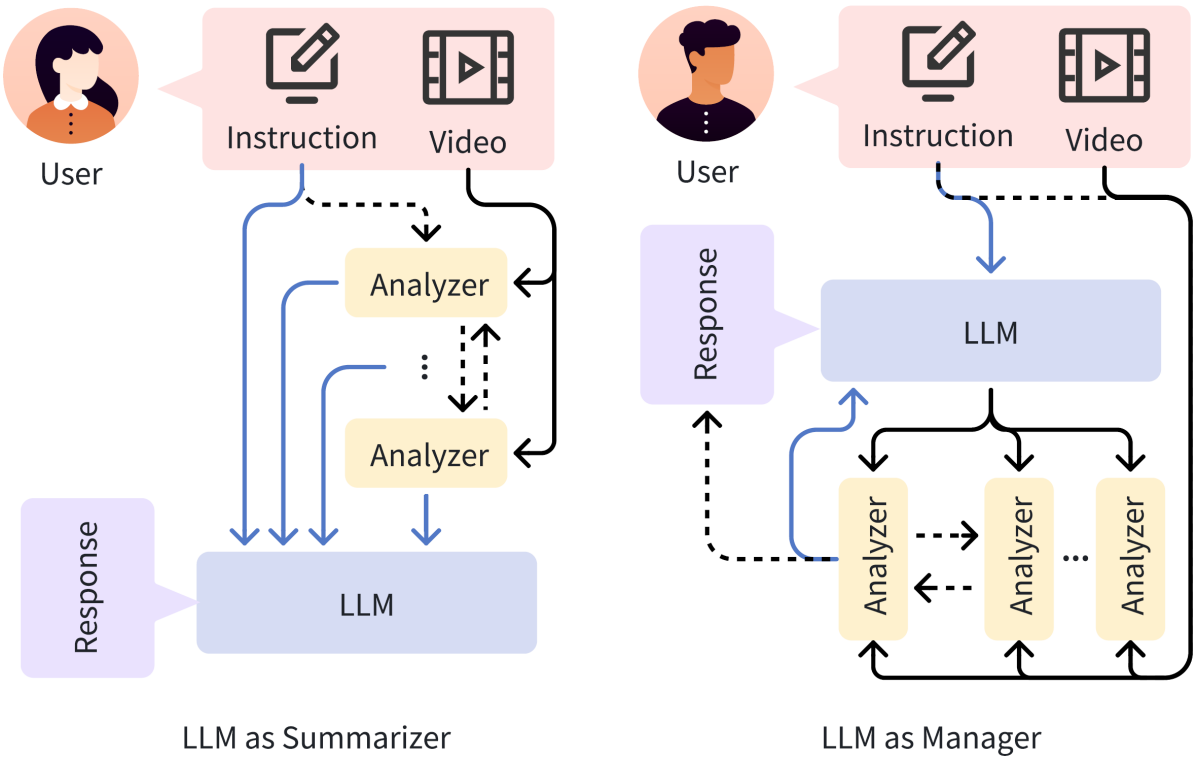

Figure 2: Illustration of Video Analyzer + LLM framework (Source: [3])

Two functional variants of this architecture are commonly employed:

1. LLM as Summarizer: in this arrangement, the LLM passively receives the output from the video analyzer and produces natural language summaries, captions, or responses. The information flow is unidirectional (i.e., Video → Analyzer → LLM), and the LLM does not influence the video processing pipeline. Notable examples include LaViLa, VAST, LLoVi, Video ReCap, Grounding-Prompter, and AntGPT [41].
2. LLM as Manager: this form assigns the LLM an active function, whereby it provides directives, oversees various analytical instruments, and participates in iterative exchanges to accomplish intricate tasks. The LLM operates as a sophisticated coordinator of perceptual modules. Notable systems in this area include ViperGPT, HuggingGPT, VideoAgent, SCHEMA, VideoTree, GPTSee, and AssistGPT [42].

This architecture is especially appealing due to its training-free and modular structure, facilitating swift prototyping and deployment of video comprehension systems with readily available LLMs and vision tools. It constitutes a fundamental pattern in the architecture of numerous contemporary Vid-LLM systems.

4.2. Video Embedder + LLM

The Video Embedder combined with the LLM architecture presents a more cohesive approach to video comprehension by embedding video content into a continuous feature space through a visual encoder (or embedder), and directly supplying these dense representations to a Large Language

Model (LLM) (Fig. 3). Unlike the Video Analyzer + LLM framework, which depends on textual intermediate outputs, this design permits the LLM to process raw visual data in embedded form, facilitating more detailed multimodal reasoning [43].

In this design, the video embedder, which may include a 3D CNN, a Transformer-based encoder, or a vision-language pretrained model, analyzes the video input V to produce a series of latent embeddings $\{v_1, v_2, ..., v_T\}$, which are then aligned with the token space of the LLM. A bridging modality mechanism, such as a linear projection, adapter module, or cross-attention layer, is often used to transform visual embeddings into a format compatible with the input space of the LLM [44].

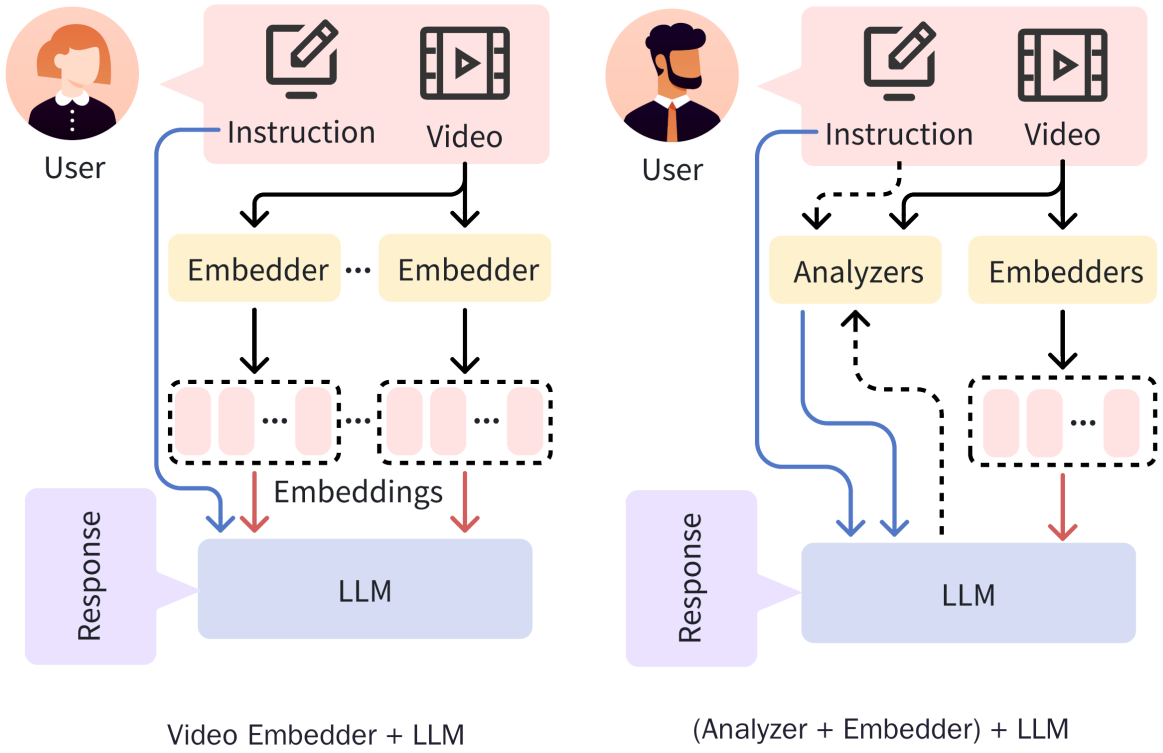

Figure 3: Illustration of Video Embedder + LLM framework and (Analyzer + Embedder) + LLM framework (Source: [3])

This method facilitates comprehensive training and enables the LLM to directly engage with visual representations, positioning it effectively for tasks that necessitate temporal alignment, spatial grounding, or multimodal reasoning across video and language inputs [45].

Training these models necessitates extensive multimodal data and meticulous design of the fusion mechanisms to guarantee stable integration of visual and textual features. Prominent instances of this architecture encompass:

1. BLIP-2, MiniGPT4, and LLaVA (adapted for video with visual embedding extensions) [46].
2. Video-ChatGPT, Video-LLaVA, Video-Chat, and MM-VID (which incorporate visual embedding adapters) [47].
3. SEED, Video-LLaMA, and mPLUG-Owl (leveraging pretrained vision-language encoders and LLMs for multimodal interaction) [48].

The Video Embedder + LLM architecture provides a cohesive multimodal interface between vision and language, facilitating enhanced interaction across modalities. Nonetheless, it often requires task specific adjustment and is susceptible to the quality of visual embeddings and their alignment with linguistic representations. This methodology signifies progress in achieving coherent video-language integration, facilitating a diverse array of downstream tasks like video captioning, question answering, and temporal localization [49].

4.3. (Analyzer + Embedder) + LLM

The design of the (Analyzer + Embedder) + LLM integrates the optimal elements of both textual and visual feature pathways by using a video analyzer in conjunction with a video embedder. The outcomes of both are further processed using a Large Language Model (LLM) (Fig. 3). This combined design seeks to use the advantages of both organized symbolic information and intricate visual imagery to enhance video comprehension and adaptability [50].

In this configuration, the video analyzer produces results that are comprehensible and often interpretable by people. These outputs include action labels, subtitles, and temporal annotations that encapsulate the video material coherently. The video embedder converts the video into a sequence of visual embeddings that capture intricate spatial and temporal details. Token union, dual-stream focus, and multimodal adapters are examples of modality fusion methods used to integrate both symbolic and visual streams into the LLM [51].

This design facilitates the LLM's execution of complex multimodal cognitive tasks by integrating advanced symbolic concepts with fundamental visual attributes. It is effective in scenarios when either symbolic or embedded information alone is insufficient, such as when simultaneous visual grounding and semantic summarization are required. Examples of systems using this design include:

1. Video-LLaVA, Video-Chat, and MM-ReAct (which combine dense vision features with analyzergenerated text) [52].
2. GPT4Tools, MM-ReAct, and MM-Vid (that enable dynamic tool use and feature fusion based on LLM-directed instructions) [53].
3. Video-ChatGPT, which leverages both vision encoders and captioning modules for multimodal dialogue and reasoning [54].

This hybrid model architecture makes it easier to be flexible and understand, and it also lets LLMdriven control dynamically organize visual and symbolic clues. However, it makes system design more complicated since the analyzer, embedder, and LLM inputs need to be carefully synchronized and aligned [55].

The (Analyzer + Embedder) + LLM model is a potential step toward creating video-language models that can do a wide range of jobs by combining different types of data and tools in real time.

5. Vid-LLMs applications

Adding Large Language Models (LLMs) to video comprehension systems has opened up new possibilities for a wide range of real-world uses. These models, especially when used with visual encoders or analytic modules, are quite flexible and may be used for tasks that require multimodal thinking, semantic comprehension, and interacting with video information in a way that is similar to how humans do it. We talk about some of the most important areas where Video-Language Models (Vid-LLMs) have had a big effect or have a lot of promise in this part.

1. Summarizing and captioning videos involves distilling content and providing textual representation for accessibility and comprehension. The automatic generation of natural language descriptions for video content represents a key application of Vid-LLMs. These systems can generate concise summaries or detailed captions by analyzing dynamic scenes and relating them to linguistic semantics. These models can generate captions that are contextually and temporally aware, as well as highly meaningful, due to the reasoning capabilities of large language models (LLMs). They can capture purpose, emotion, and narrative structure, alongside object or action recognition. This capability is essential for activities such as content generation, material categorization, and support for individuals with disabilities [56].

2. Video Question Answering (Video QA.Vid-LLMs) lets people ask natural language questions about the video material, which makes it possible to interactively interpret videos. The model takes in both visual and spoken inputs and gives correct answers. This job needs not only object or event identification, but also spatio-temporal thinking, comprehension of causation, and even fundamental knowledge. Vid-LLMs show a lot of promise in this area, even when there are just a few examples or none at all. This is because they can follow instructions and respond to prompts, which they got from LLMs [57].
3. Temporal and Spatial Localization. A significant application area pertains to identifying particular moments, actions, or objects within video streams. This encompasses temporal action localization, moment retrieval, and referring object grounding. In these contexts, Vid-LLMs can effectively associate language-based prompts (e.g., "when does the person start running?") with specific temporal segments and spatial areas of interest within the video. This functionality is crucial for applications such as video indexing, surveillance, sports analysis, and the comprehension of instructional content [57].
4. Multimodal Video Dialogues. The rise of multimodal chat systems has led to the integration of Vid-LLMs in interactive dialogue interfaces that encompass video comprehension. These systems facilitate natural conversations that include follow-up questions, temporal references, and iterative reasoning related to video content. This approach is especially beneficial in the realms of educational technology, customer support automation, and interactive storytelling, as comprehending video context is essential for producing relevant and coherent dialogue [58].
5. Using Tools and Video-Based Agents. Recent Vid-LLM designs improve their usefulness by acting as independent agents that can understand video input, think logically, and utilize other tools or APIs as needed. These agents can work with long videos, make guesses, assess their work using tools like object detectors, trackers, and summarizers, and provide multi-step answers. Because they act like agents, they are excellent at challenging occupations that require making choices, like as robotics, autonomous monitoring, or interpreting scientific movies.
6. Retrieval and recommendation across modes. Vid-LLMs are being utilized more and more in systems that enable you search for videos and text and vice versa, where it is vital for the semantics of the two forms of material to line up. By placing both types of data into a shared latent space, these systems make it easier to identify relevant content based on natural language descriptions or the other way around. This tool is highly helpful for searching video databases, recommending material, and managing digital assets.
7. Monitoring, security, and compliance. In high-stakes situations like surveillance or forensic analysis, Vid-LLMs could help find events, spot strange behavior, or check for conformity with regulatory standards. Their ability to assess and express visual information in plain English enables transparent and verifiable decision-making, particularly advantageous in legal, security, and auditing contexts [59].

The application range of Vid-LLMs encompasses descriptive, analytical, interactive, and operational domains, facilitated by their ability to merge vision and language within a cohesive, human-centered framework. As the discipline advances, we foresee wider use of these models in domains necessitating explainability, multi-turn interaction, and dynamic management of multimodal inputs. Furthermore, forthcoming advancements in fine-tuning methodologies, long-context modeling, and the incorporation of domain-specific tools will likely enhance their applicability significantly.

6. Future research opportunities

As Large Language Models (LLMs) become more common in video understanding systems, a number of important research areas and problems arise that are likely to impact the future

generation of multimodal intelligence. Current Vid-LLMs have shown great promise in video reasoning, captioning, and interactivity, but they still have big problems with scalability, accuracy, interpretability, and generalization across domains.

1. *Video modeling with a long context*. One of the most important problems is figuring out how to analyze long-form films well. When working with long temporal sequences, current models generally have trouble with memory and processing limitations. Future research should investigate more effective temporal compression methods, hierarchical modeling, and sparse attention processes that facilitate the representation and retrieval of relevant parts over extended periods while preserving essential context [39].
2. *Learning to Tokenize and Represent Video*. Video data does not have a widely used and effective way to break it down into tokens, unlike photos or text. It's important to come up with better ways to turn raw video into symbolic or discrete representations that are both useful and easy for computers to work with. Improvements in video-language tokenizers, separate visual vocabularies, and multimodal pretraining goals will be important for making video work better with LLM structures [60].
3. *Reasoning that is based on facts and can be explained*. Future Vid-LLMs must be able to do grounded reasoning, which means that predictions must be clearly connected to particular visual or temporal data in the input video. This is necessary for real-world applications to be trustworthy and open. Developing methods that produce explainable and verifiable outputs – for instance, via textual rationales, highlighted video frames, or traceable inference paths – will be vital for adoption in sensitive domains such as healthcare, legal analysis, and autonomous systems [61].
4. *Dynamic interaction and adding tools*. The trend toward LLM-driven agents is expected to continue. Vid-LLMs will be able to use outside resources to help them see, track, summarize, and find things. Future systems could be better at planning and reasoning if they have better memory, self-correction, and adaptive tool invocation mechanisms. Research into multi-agent collaboration, where different expert models cooperate via LLM coordination, is also a promising direction [62].
5. *Grounded and explainable reasoning*. Future Vid-LLMs must demonstrate grounded reasoning to ensure trust and transparency in real-world applications, linking predictions explicitly to specific visual or temporal evidence in the input video. Creating methods that yield explainable and verifiable outputs – such as textual rationales, highlighted video frames, or traceable inference paths – will be essential for implementation in sensitive fields like healthcare, legal analysis, and autonomous systems [61].
6. *Dynamic interaction and tool enhancement*. The trend of LLM-driven agents is expected to persist, with Vid-LLMs enhanced by external tools for perception, tracking, summarization, and retrieval. Future systems could improve through advanced planning and reasoning abilities, encompassing memory, self-correction, and adaptive strategies for tool invocation. Investigating multi-agent collaboration, in which various expert models interact through LLM coordination, represents a promising avenue of research [62].

7. Conclusion

In recent years, the integration of video analysis with extensive language modeling has significantly transformed the field of multimodal artificial intelligence. Video-Language Models (Vid-LLMs) represent an advanced type of system capable of performing complex reasoning, engaging in interactive dialogue, and demonstrating semantic understanding of video content. Their functionality is achieved through the integration of Large Language Models (LLMs) with vision encoders and perceptual tools.

This survey has given a full picture of how Vid-LLMs have changed over time, how they are built, how they are used, and what problems they face. We put current methods into three main

architectural paradigms: Video Analyzer + LLM, Video Embedder + LLM, and (Analyzer + Embedder) + LLM. We did this by showing how they work, how they are designed, and what models they are based on. We also looked at a lot of real-world uses, such as captioning, answering questions, temporal localization, multimodal conversation, retrieval, and agentic reasoning.

Even though modern Vid-LLMs are quite powerful, they are still in the early stages of development. Their performance is generally limited by problems with video tokenization, long-context modeling, multimodal alignment, and explainability. The research community still has a lot of work to do on topics like scalability, domain robustness, and standardizing evaluations.

Integrating LLMs into video understanding represents a significant advancement in the development of general-purpose, instruction-driven, and human-aligned multimodal AI systems. With advancements in modeling architectures, pretraining techniques, and system-level design, Vid-LLMs are poised to play a significant role in the development of intelligent systems capable of seamlessly interacting with their visual environments in the future.

This study lays the groundwork for understanding the current landscape of Vid-LLMs and offers a structured approach for future research initiatives aimed at improving the capabilities, efficiency, and reliability of video-language models.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, S. Hoi, InstructBLIP: towards general-purpose vision-language models with instruction tuning, arXiv preprint arXiv:2305.06500 (2023). doi:10.48550/arXiv.2305.06500.

[2] D. M. Argaw, S. Yoon, F. C. Heilbron, H. Deilamsalehy, T. Bui, Z. Wang, F. Dernoncourt, J. S. Chung, Scaling up video summarization pretraining with large language models, in: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR '2024, IEEE, New York, NY, 2024, pp. 8332–8341. doi:10.1109/CVPR52733.2024.00796.

[3] Y. Tang, J. Bi, S. Xu, L. Song, S. Liang, T. Wang, D. Zhang, J. An, J. Lin, R. Zhu, A. Vosoughi, C. Huang, Z. Zhang, P. Liu, M. Feng, F. Zheng, J. Zhang, P. Luo, J. Luo, C. Xu, Video understanding with large language models: a survey, arXiv preprint arXiv:2312.17432 (2025). doi:10.48550/arXiv.2312.17432.

[4] OpenAI, GPT-4 technical report, arXiv preprint arXiv:2303.08774 (2024). doi:10.48550/arXiv.2303.08774.

[5] H. Qingnan, C. Xiaodong, Z. Meixin, W. Xiangqing, Video content understanding based on spatio-temporal feature extraction and pruning network, in: Proceedings of the 2023 4th International Symposium on Computer Engineering and Intelligent Communications, ISCEIC '2023, IEEE, New York, NY, 2023, pp. 493–497. doi:10.1109/ISCEIC59030.2023.10271098.

[6] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023). doi:10.48550/arXiv.2302.13971.

[7] T. Yuan, X. Zhang, B. Liu, K. Liu, J. Jin, Z. Jiao, Surveillance video-and-language understanding: from small to large multimodal models, IEEE Trans. Circuits Syst. Video Technol. 35 (2025) 300–314. doi:10.1109/TCSVT.2024.3462433.

[8] T. Lindeberg, Scale invariant feature transform, Scholarpedia 7 (2012) 10491. doi:10.4249/scholarpedia.10491.

[9] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Comput. Vis. Image Underst. 110 (2008) 346–359. doi:10.1016/j.cviu.2007.09.014.

[10] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'05, IEEE, New York, NY, 2005, pp. 886–893. doi:10.1109/CVPR.2005.177.
[11] H. Wang, C. Schmid, Action recognition with improved trajectories, in: Proceedings of the 2013 IEEE International Conference on Computer Vision, ICCV '2013, IEEE, New York, NY, 2013, pp. 3551–3558. doi:10.1109/ICCV.2013.441.
[12] X. Liu, T. Cheng, Video-based face recognition using adaptive hidden Markov models, in: Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '2003, IEEE, New York, NY, 2003, pp. I–I. doi:10.1109/CVPR.2003.1211373.
[13] H. Sidenbladh, Detecting human motion with support vector machines, in: Proceedings of the Pattern Recognition, International Conference on, ICPR '2004, IEEE, New York, NY, 2004, pp. 188–191. doi:10.1109/ICPR.2004.1334092.
[14] A. Mittal, S. Gupta, Automatic content-based retrieval and semantic classification of video content, Int. J. Digit. Libr. 6 (2006) 30–38. doi:10.1007/s00799-005-0119-y.
[15] A. B. Chan, N. Vasconcelos, Modeling, clustering, and segmenting video with mixtures of dynamic textures, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2008) 909–926. doi:10.1109/TPAMI.2007.70738.
[16] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '2014, IEEE, New York, NY, 2014, pp. 1725–1732. doi:10.1109/CVPR.2014.223.
[17] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 221–231. doi:10.1109/TPAMI.2012.59.
[18] F. Salem, Recurrent neural networks: from simple to gated architectures, 2022. doi:10.1007/978-3-030-89929-5.
[19] B. Lindemann, T. Müller, H. Vietz, N. Jazdi, M. Weyrich, A survey on long short-term memory networks for time series prediction, Procedia CIRP 99 (2021) 650–655. doi:10.1016/j.procir.2021.03.088.
[20] G. Yang, Y. Yang, Z. Lu, J. Yang, D. Liu, C. Zhou, Z. Fan, STA-TSN: spatial–temporal attention temporal segment network for action recognition in video, PLoS ONE 17 (2022) 1–19. doi:10.1371/journal.pone.0265115.
[21] M. Sekma, M. Mejdoub, C. Ben Amar, Human action recognition based on multi-layer Fisher vector encoding method, Pattern Recognit. Lett. 65 (2015) 37–43. doi:10.1016/j.patrec.2015.06.029.
[22] A. Diba, V. Sharma, L. Van Gool, Deep temporal linear encoding networks, in: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '2017, IEEE, New York, NY, 2017, pp. 1541–1550. doi:10.1109/CVPR.2017.168.
[23] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the Kinetics dataset, in: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '2017, IEEE, New York, NY, 2017, pp. 4724–4733. doi:10.1109/CVPR.2017.502.
[24] A. Gabriel, S. Cosar, N. Bellotto, P. Baxter, A dataset for action recognition in the wild, in: Annual Conference Towards Autonomous Robotic Systems, Springer, London, UK, 2019, pp. 362–374. doi:10.1007/978-3-030-23807-0_30.
[25] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: a large video database for human motion recognition, in: 2011 International Conference on Computer Vision (ICCV), 2011, pp. 2556–2563. doi:10.1109/ICCV.2011.6126543.
[26] K. Hara, H. Kataoka, Y. Satoh, Learning spatio-temporal features with 3D residual networks for action recognition, arXiv preprint arXiv:1708.07632 (2017). doi:10.48550/arXiv.1708.07632.
[27] Q. Yu, Y. Li, J. Mei, Y. Zhou, A. L. Yuille, CAKES: channel-wise automatic kernel shrinking for efficient 3D networks, arXiv preprint arXiv:2003.12798 (2020). doi:10.48550/arXiv.2003.12798.
[28] J. Lao, W. Hong, X. Guo, Y. Zhang, J. Wang, J. Chen, W. Chu, Simultaneously short- and long-term temporal modeling for semi-supervised video semantic segmentation, in: Proceedings of

the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR '2023, IEEE, New York, NY, 2023, pp. 14763–14772. doi:10.1109/CVPR52729.2023.01418.
[29] C. Feichtenhofer, X3D: expanding architectures for efficient video recognition, arXiv preprint arXiv:2004.04730 (2020). doi:10.48550/arXiv.2004.04730.
[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16×16 words: transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2021). doi:10.48550/arXiv.2010.11929.
[31] G. Bertasius, H. Wang, L. Torresani, Is space–time attention all you need for video understanding?, arXiv preprint arXiv:2102.05095 (2021). doi:10.48550/arXiv.2102.05095.
[32] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid, ViViT: a video vision transformer, arXiv preprint arXiv:2103.15691 (2021). doi:10.48550/arXiv.2103.15691.
[33] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, C. Feichtenhofer, Multiscale vision transformers, arXiv preprint arXiv:2104.11227 (2021). doi:10.48550/arXiv.2104.11227.
[34] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in BERTology: what we know about how BERT works, Trans. Assoc. Comput. Linguist. 8 (2020) 842–866. doi:10.1162/tacl_a_00349.
[35] L. Zhu, Y. Yang, ActBERT: learning global–local video–text representations, in: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR '2020, IEEE, New York, NY, 2020, pp. 8743–8752. doi:10.1109/CVPR42600.2020.00877.
[36] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, C. Feichtenhofer, Masked feature prediction for self-supervised visual pre-training, arXiv preprint arXiv:2112.09133 (2023). doi:10.48550/arXiv.2112.09133.
[37] T. Zhang, C. Xu, G. Zhu, S. Liu, H. Lu, A generic framework for video annotation via semi-supervised learning, IEEE Trans. Multimedia 14 (2012) 1206–1219. doi:10.1109/TMM.2012.2191944.
[38] L. Song, G. Yin, B. Liu, Y. Zhang, N. Yu, FSFT-Net: face transfer video generation with few-shot views, in: Proceeding of the 2021 IEEE International Conference on Image Processing, ICIP' 2021, IEEE, New York, NY, 2021, pp. 3582–3586. doi:10.1109/ICIP42928.2021.9506512.
[39] K. Chen, M. Hu, An automatic video tag extraction method based on large language model text content parsing, in: 2024 7th International Conference on Data Science and Information Technology (DSIT), 2024, pp. 1–4. doi:10.1109/DSIT61374.2024.10881284.
[40] W. Wen, Y. Wang, N. Birkbeck, B. Adsumilli, An ensemble approach to short-form video quality assessment using multimodal LLM, in: Proceeding of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '2025, IEEE, New York, NY, 2025, pp. 1–5. doi:10.1109/ICASSP49660.2025.10888524.
[41] Q. Zhao, S. Wang, C. Zhang, C. Fu, M. Q. Do, N. Agarwal, K. Lee, C. Sun, AntGPT: can large language models help long-term action anticipation from videos?, arXiv preprint arXiv:2307.16368 (2024). doi:10.48550/arXiv.2307.16368.
[42] D. Surís, S. Menon, C. Vondrick, ViperGPT: visual inference via Python execution for reasoning, arXiv preprint arXiv:2303.08128 (2023). doi:10.48550/arXiv.2303.08128.
[43] C. Hori, M. Kambara, K. Sugiura, K. Ota, S. Khurana, S. Jain, R. Corcodel, D. Jha, D. Romeres, J. Le Roux, Interactive robot action replanning using multimodal LLM trained from human demonstration videos, in: Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '2025, IEEE, New York, NY, 2025, pp. 1–5. doi:10.1109/ICASSP49660.2025.10887717.
[44] R. Liu, C. Li, Y. Ge, T. H. Li, Y. Shan, G. Li, BT-Adapter: video conversation is feasible without video instruction tuning, in: Proceeding of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR '2024, IEEE, New York, NY, 2024, pp. 13658–13667. doi:10.1109/CVPR52733.2024.01296.
[45] M. Bain, A. Nagrani, G. Varol, A. Zisserman, Frozen in time: a joint video and image encoder for end-to-end retrieval, in: Proceeding of the 2021 IEEE/CVF International Conference on

Computer Vision, ICCV ' 2021, IEEE, New York, NY, 2021, pp. 1708–1718. doi:10.1109/ICCV48922.2021.00175.

[46] K. Ataallah, X. Shen, E. Abdelrahman, E. Sleiman, D. Zhu, J. Ding, M. Elhoseiny, MiniGPT4-Video: advancing multimodal LLMs for video understanding with interleaved visual–textual tokens, arXiv preprint arXiv:2404.03413 (2024). doi:10.48550/arXiv.2404.03413.

[47] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, Y. Qiao, VideoChat: chat-centric video understanding, arXiv preprint arXiv:2305.06355 (2024). doi:10.48550/arXiv.2305.06355.

[48] H. Zhang, X. Li, L. Bing, Video-LLaMA: an instruction-tuned audio-visual language model for video understanding, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Singapore, 2023, pp. 543–553. doi:10.18653/v1/2023.emnlp-demo.49.

[49] M. Yilmazer, M. Karakose, LLM-based video analytics test scenario generation in smart cities, in: 2025 29th International Conference on Information Technology (IT), 2025, pp. 1–4. doi:10.1109/IT64745.2025.10930297.

[50] R. Ma, Y. Yang, Z. Liu, J. Zhang, M. Li, J. Huang, G. Luo, VerilogReader: LLM-aided hardware test generation, in: Proceeding of the 2024 IEEE LLM Aided Design Workshop, LAD '2024, IEEE, New York, NY, 2024, pp. 1–5. doi:10.1109/LAD62341.2024.10691801.

[51] Q. Guo, J. Cao, X. Xie, S. Liu, X. Li, B. Chen, X. Peng, Exploring the potential of ChatGPT in automated code refinement: an empirical study, in: Proceedings of the IEEE/ACM 46th International Conference on Software Engineering, ICSE '24, Association for Computing Machinery, New York, NY, 2024, pp. 1–13. doi:10.1145/3597503.3623306.

[52] S. Munasinghe, R. Thushara, M. Maaz, H. A. Rasheed, S. Khan, M. Shah, F. Khan, PG-VideoLLaVA: pixel grounding large video–language models, arXiv preprint arXiv:2311.13435 (2023). URL: doi:10.48550/arXiv.2311.13435.

[53] K. Lin, F. Ahmed, L. Li, C.-C. Lin, E. Azarnasab, Z. Yang, J. Wang, L. Liang, Z. Liu, Y. Lu, C. Liu, L. Wang, MM-VID: advancing video understanding with GPT-4V(ision), arXiv preprint arXiv:2310.19773 (2023). doi:10.48550/arXiv.2310.19773.

[54] S. Wang, Q. Zhao, M. Q. Do, N. Agarwal, K. Lee, C. Sun, VAMOS: versatile action models for video understanding, arXiv preprint arXiv:2311.13627 (2024). doi:10.48550/arXiv.2311.13627.

[55] B. Huang, X. Wang, H. Chen, Z. Song, W. Zhu, VTimeLLM: empower LLM to grasp video moments, in: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR '2024, IEEE, New York, NY, 2024, pp. 14271–14280. doi:10.1109/CVPR52733.2024.01353.

[56] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, B. Russell, Localizing moments in video with natural language, in: Proceeding of the 2017 IEEE International Conference on Computer Vision, ICCV '2017, IEEE, New York, NY, 2017, pp. 5804–5813. doi:10.1109/ICCV.2017.618.

[57] H. Wang, K. Hu, L. Gao, DocVideoQA: towards comprehensive understanding of document-centric videos through question answering, in: Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '2025, IEEE, New York, NY, 2025, pp. 1–5. doi:10.1109/ICASSP49660.2025.10887668.

[58] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, S. Lacoste-Julien, Unsupervised learning from narrated instruction videos, in: Proceeding of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '2016, IEEE, New York, NY, 2016, pp. 4575–4583. doi:10.1109/CVPR.2016.495.

[59] S. Jin, C. Tang, Y. Li, Research on logical understanding of video surveillance systems based on knowledge graphs, in: 2023 4th International Conference on Computer, Big Data and Artificial Intelligence (ICCBD+AI), 2023, pp. 617–622. doi:10.1109/ICCBD-AI62252.2023.00113.

[60] A. Senthilselvi, R. Prawin, V. Harshit, R. Santhosh Kumar, S. Senthil Pandi, Abstractive summarization of YouTube videos using Lamini-FLAN-T5 LLM, in: 2024 Second International

Conference on Advances in Information Technology (ICAIT), vol. 1, 2024, pp. 1–5. doi:10.1109/ICAIT61638.2024.10690747.

[61] C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. Liao, T. Gao, E. Li, K. Tang, Z. Cao, T. Zhou, A. Liu, X. Yan, S. Mei, J. Cao, Z. Wang, C. Zheng, A survey on multimodal large language models for autonomous driving, in: Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACVW '2024, IEEE, New York, NY, 2024, pp. 958–979. doi:10.1109/WACVW60836.2024.00106.

[62] Y. Chen, J. Arkin, Y. Zhang, N. Roy, C. Fan, Scalable multi-robot collaboration with large language models: centralized or decentralized systems?, in: Proceedings, of the 2024 IEEE International Conference on Robotics and Automation, ICRA '2024, IEEE, New York, NY, 2024, pp. 4311–4317. doi:10.1109/ICRA57147.2024.10610676.

[63] H. Qi, L. Dai, W. Chen, Z. Jia, X. Lu, Performance characterization of large language models on high-speed interconnects, in: Proceeding of the 2023 IEEE Symposium on High-Performance Interconnects, HOTI '2023, IEEE, New York, NY, 2023, pp. 53–60. doi:10.1109/HOTI59126.2023.00022.