applied sciences

MDPI

Review

Multimodal Video Understanding: A Capability-Based Survey of Alignment, Expression, and Reasoning

Rongyong Zhao [1], Da Pu [1,*], Cuiling Li [2], Zeyi Zhang [1], Xingzhu Peng [1] and Yunlong Ma [1]

[1] CIMS Research Center, Tongji University, 4800 Cao'an Road, Shanghai 201804, China; zhaorongyong@tongji.edu.cn (R.Z.); 2531846@tongji.edu.cn (Z.Z.); 2531844@tongji.edu.cn (X.P.); evanma@tongji.edu.cn (Y.M.)

[2] The College of Electronic and Information Engineering, Tongji University, 4800 Cao'an Road, Shanghai 201804, China; licuiling@tongji.edu.cn

* Correspondence: puda@tongji.edu.cn

Abstract

Multimodal video understanding (MVU) has emerged as a fast-growing research frontier, driven by major advances in video-language pre-training and large multimodal models over the past decade. MVU aims to synergistically integrate visual, audio and textual modalities to interpret complex video semantics, supporting widespread downstream tasks including cross-modal retrieval, dense captioning, video question answering, event analysis and intelligent assistance. Despite the rapid proliferation of specialized MVU models, the field still lacks a unified capability-centric framework to systematically clarify the hierarchical competency architecture and evolutionary trajectory of state-of-the-art approaches. To address this issue, this paper presents a structured, comprehensive survey of the latest MVU progress, establishing a novel three-tier taxonomy that categorizes existing studies into cross-modal alignment, multi-granularity semantic expression and multimodal reasoning. Along this pipeline, we further systematically synthesize core modality fusion strategies, mainstream benchmark datasets and standardized evaluation protocols. Through a fine-grained analysis of representative published results, we highlight the critical impact of inconsistent evaluation settings, cross-experiment comparability bottlenecks and inherent methodological trade-offs between performance and efficiency. Finally, we identify and dissect three key open challenges: ultra-long video scalability, performance degradation from modality noise and missing data, and factual reliability risks in generative MVU systems. This capability-oriented systematic reference clarifies the methodological evolution logic of MVU, and provides actionable guidance for developing next-generation robust, high-performance multimodal video understanding systems.

Keywords: multimodal video understanding; large multimodal models; text–video retrieval; video captioning; video question answering

Check for updates

Academic Editor: João M. F. Rodrigues

Received: 10 June 2026
Revised: 9 July 2026
Accepted: 14 July 2026
Published: 16 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The proliferation of the mobile internet and social media has led to an exponential increase in short video content, which is characterized by frequent uploads, rapid editing, and parallel information delivery across multiple channels. In this context, manual methods for retrieving, filtering, and semantically summarizing large-scale video content are inefficient and cannot meet the requirements of real-time and large-scale applications. Therefore, it is essential to develop models that can automatically extract high-level semantics from videos and support a broad range of video understanding tasks.

Video understanding spans multiple levels, from perception to cognition. Early studies mainly focused on fundamental tasks such as action recognition, which aims to determine "what happened in the video" [1,2]. Subsequent research extended this line of work to temporal boundary modeling and sequence localization, further addressing "when it happened" and "in which segment of the video" [3]. In real-world applications, category recognition and localization alone are often insufficient for satisfying complex information needs. Consequently, recent research has increasingly emphasized semantic generation and high-level reasoning, including tasks such as video captioning and video question answering [4,5]. These directions are not mutually exclusive; rather, they have developed in parallel within a unified framework of video representation learning and temporal modeling. In recent years, they have also become closely integrated with large-scale pretraining and generative modeling paradigms.

Unlike static images or plain text, videos inherently contain a temporal dimension and are often accompanied by multiple sources of information, including speech, background music, on-screen text, bullet-screen comments, and captions. As a result, video semantic representations strongly depend on cross-modal complementarity and alignment. Traditional video analysis methods primarily rely on visual signals, using convolutional backbones and temporal modeling modules to extract spatio-temporal features for classification and localization [2,3,6]. However, in short video scenarios, semantic cues are often distributed across non-visual channels such as audio content, subtitle text, and music-induced affective cues. Relying solely on visual information can therefore lead to systematic gaps in understanding event motivations, dialogue content, and emotional tendencies.

1.1. Related Work

Existing studies have reviewed multimodal video understanding and related topics from different perspectives. As summarized in Table 1, several studies focus on general multimodal learning or video-language modeling. Xu et al. [7] reviewed multimodal learning with Transformers, covering representative architectures, pretraining paradigms, and task-specific applications. Nguyen et al. [8] surveyed video-language understanding from the perspectives of model architecture, training strategies, and data construction. Tang et al. [9] further examined video understanding with large language models, focusing on recent Vid-LLM architectures, training strategies, datasets, and evaluation protocols. These surveys provide valuable methodological or large-model-oriented perspectives. However, they do not explicitly organize multimodal video understanding from a functional capability perspective that spans cross-modal matching, semantic generation, and evidence-based reasoning.

Table 1. Comparison of representative surveys related to multimodal video understanding.

Survey	Venue/ Year	Mod.	Primary Scope	Main Coverage
Xu et al. [7]	TPAMI 2023	V + A + T	Multimodal Transformers	Transformer-based multimodal modeling, pretraining, and applications.
Zhu et al. [10]	IJMIR 2023	V + T	Video-text retrieval	Retrieval models, cross-modal matching, benchmarks, and metrics.
Zhang et al. [11]	TPAMI 2023	V + T	Temporal sentence grounding	Language-based temporal localization and video moment retrieval.
Agrawal et al. [12]	CSET 2023	V + T	Visual question answering	VQA methods, tools, datasets, and evaluation settings.
Abdar et al. [4]	TPAMI 2024	V + A + T	Video captioning	Captioning models, datasets, metrics, and applications.
Nguyen et al. [8]	ACL Findings 2024	V + T	Video-language understanding	Model architectures, training strategies, data construction, and tasks.

Table 1. *Cont.*

Survey	Venue/ Year	Mod.	Primary Scope	Main Coverage
Jeshmol and Kovoor [5]	JVCIR 2024	V + T	Video question answering	VideoQA methods, datasets, evaluation settings, and state-of-the-art results.
Qasim et al. [13]	ACM CSUR 2025	V + T	Dense video captioning	Feature extraction, event localization, caption generation, and protocols.
Tang et al. [9]	TCSVT 2025	V + A + T	Video understanding with LLMs	Vid-LLM architectures, training strategies, datasets, and evaluation.

Note: V, A, and T denote vision, audio, and text, respectively.

Other surveys focus on specific task families. For alignment-oriented tasks, Zhu et al. [10] reviewed video-text retrieval, while Zhang et al. [11] focused on temporal sentence grounding. For expression-oriented tasks, Abdar et al. [4] reviewed deep learning methods for video captioning, and Qasim et al. [13] provided a dedicated survey of dense video captioning techniques, datasets, and evaluation protocols. For reasoning-oriented tasks, Agrawal et al. [12] summarized methods and datasets for visual question answering, while Jeshmol and Kovoor [5] focused specifically on video question answering. Although these studies provide detailed analyses of individual subfields, they typically treat retrieval, grounding, captioning, and question answering as separate research directions. In contrast, this survey proposes a capability-oriented framework that categorizes multimodal video understanding into alignment, expression, and reasoning. From this unified perspective, this survey analyzes representative tasks, models, datasets, metrics, benchmark comparability, and engineering considerations, thereby providing a cross-task perspective that complements existing surveys.

1.2. Literature Search and Selection Strategy

In preparing Section 3, the literature search was conducted primarily in Web of Science and IEEE Xplore, with arXiv used as a supplementary source for recent preprints and conference-related works. The search focused on studies published or publicly available between 2021 and 2025. The main search terms included broad queries such as "video" and "video understanding", combined with multimodal or cross-modal terms such as "multimodal", "cross-modal", "video-language", and "multimodal video understanding". To capture task-specific studies, these terms were further combined with task keywords including "captioning", "retrieval", "matching", "question answering", and "VideoQA". Several earlier works were retained when necessary to introduce classical models, datasets, evaluation metrics, or background concepts, but they were not counted as part of the recent core corpus.

Candidate papers were first screened by reading their titles, keywords, and abstracts, and full texts were further checked when relevance was not immediately clear. For the core corpus, studies were included if they focused on multimodal video understanding and addressed at least one capability layer in the proposed framework: alignment, expression, or reasoning. Eligible studies were required to involve video together with textual, audio, speech-transcript, subtitle, OCR, or natural-language information, and to provide sufficient methodological and experimental details for analysis. Among eligible studies, priority was given to representative methods that were widely used in benchmark comparisons, reflected major technical trends, or appeared in major journals and conferences.

Studies were excluded if they focused only on static images, pure text, pure audio, or unimodal video recognition; if they did not address cross-modal video understanding; if their objectives were unrelated to alignment, expression, or reasoning; if they lacked sufficient methodological or experimental details; or if they were duplicate versions of already included studies. Based on this process, 57 representative studies published or publicly available between 2021 and 2025 were selected and organized in Section 3.

Figure 1 presents a keyword co-occurrence visualization of these 57 selected studies. The bibliographic information of the selected papers was imported into VOSviewer (version 1.6.21), and the keyword co-occurrence function was used to identify frequently associated research topics.

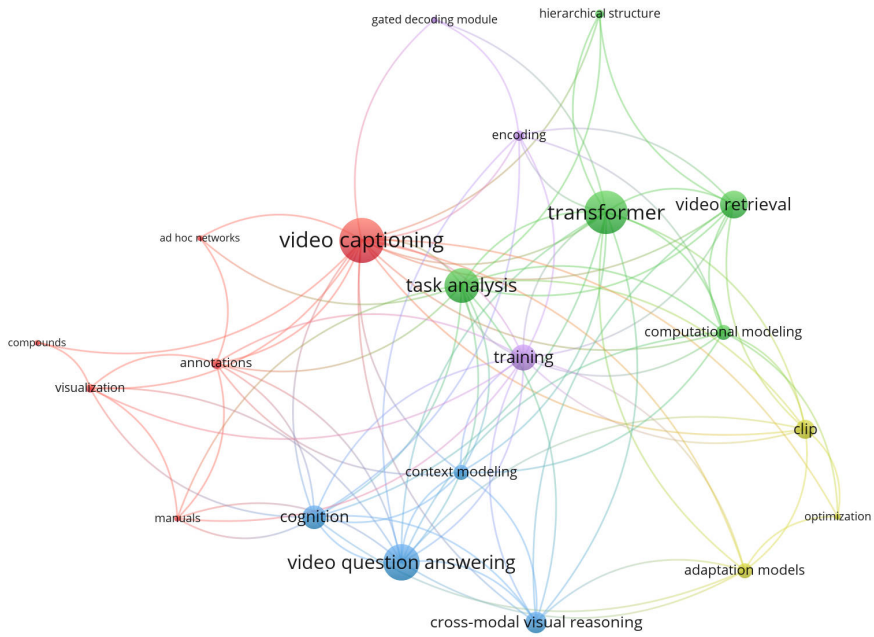

Figure 1. Keyword co-occurrence network of the selected studies on multimodal video understanding. Nodes represent keywords, node size indicates occurrence frequency, links denote keyword co-occurrence, and colors represent clusters of related research topics.

The keyword co-occurrence network shows that the field is organized around a strong central cluster formed by task analysis and training, which both have high connectivity with other terms. This indicates that current work is still largely driven by task design and model optimization rather than by a clear separation of conceptual problems. Around this center, video question answering, cross-modal visual reasoning, and cognition-related terms form a closely linked group, reflecting increasing attention to reasoning-oriented video understanding. In parallel, video captioning and video retrieval are positioned closer to transformer-related terms, suggesting that alignment and generation tasks are increasingly built on shared transformer-based architectures rather than developed as independent modeling directions. The structure of the network suggests a clear research trend toward unified transformer-based modeling. At the same time, reasoning-oriented video understanding is becoming a main research focus, while alignment and generation tasks are being gradually absorbed into a common modeling framework.

To further clarify the meaning of representativeness, the 57-paper core corpus was not intended to be an exhaustive collection of all publications on multimodal video understanding. Instead, it was constructed to provide structured coverage of the main capability layers, task families, modeling paradigms, modality combinations, and evaluation protocols that are necessary for building the taxonomy and conducting trend analysis in this survey. From a methodological perspective, the corpus covers contrastive video-language pretraining, CLIP-based transfer, cross-modal attention, hierarchical temporal modeling, memory-based modeling, multimodal large language models, and evidence-aware reasoning. Table 2 summarizes the structural coverage of the selected core corpus. This structured coverage was used as an additional criterion when organizing the selected studies. Therefore, the term "representative" in this survey refers to coverage of the major methodological and

task-oriented dimensions of multimodal video understanding, rather than to a complete enumeration of all relevant studies.

Table 2. Coverage of the core corpus.

Capability Layer	Covered Task Families	Covered Modeling Paradigms	Typical Evaluation Focus
Alignment	Video–text retrieval, video–text matching, temporal grounding, cross-modal correspondence learning	Contrastive video-language pretraining, dual-encoder retrieval, CLIP-based transfer, global–local matching, hierarchical cross-modal interaction, robust retrieval under noisy supervision	Recall@K, median rank, average precision, temporal localization accuracy
Expression	Video captioning, dense video captioning, video summarization, emotion-aware captioning, domain-specific video description	Transformer-based captioning, end-to-end dense captioning, hierarchical event organization, non-autoregressive generation, memory-based and streaming description, knowledge-enhanced generation	BLEU, METEOR, ROUGE-L, CIDEr, event coverage, semantic consistency, temporal coherence
Reasoning	Video question answering, temporal reasoning, causal reasoning, multi-hop reasoning, evidence-based long-video understanding	Spatio-temporal graph reasoning, question-guided attention, semantic decoupling, causal intervention, counterfactual reasoning, multimodal LLMs, modular and tool-assisted reasoning	Answer accuracy, causal/temporal reasoning consistency, evidence localization, robustness to shortcuts

Notably, this core corpus does not include all references cited in this survey. Earlier foundational studies on video representation learning, multimodal fusion, and vision-language pretraining are additionally cited to introduce background concepts and trace historical development. Dataset papers, metric papers, and existing survey articles are also included where necessary to support the discussion of benchmarks, evaluation protocols, and related work. Therefore, the 57-paper corpus specifically refers to recent research studies used for method taxonomy and trend analysis, rather than the complete bibliography of this survey.

All figures in this paper are original illustrations created by the authors, except Figure 1, which was generated using VOSviewer based on the selected corpus and properly attributed in its caption.

1.3. Contributions

To systematically organize research progress and technical frontiers in multimodal video understanding, the contributions of this survey are summarized as follows:

1. Unified classification framework: This survey proposes a progressive framework based on key capabilities in multimodal video understanding. This framework integrates existing methods into a unified analytical structure, thereby facilitating cross-task comparison and the identification of research gaps.
2. Technological lineage synthesis: This survey systematically reviews and compares representative approaches along the evolution of representation learning and modeling paradigms, and summarizes their key design choices and applicable conditions.
3. Resource and evaluation consolidation: This survey aggregates mainstream datasets and commonly used evaluation metrics for multimodal video understanding, discusses comparability issues across different evaluation settings, and establishes a reusable reference framework.
4. Performance and trend analysis: Based on publicly available benchmark results, this survey summarizes the performance evolution of representative models and further

identifies major challenges and future directions in this field, including long-video modeling, cross-modal alignment, and generative reasoning.

1.4. Structure of the Survey

The remainder of this survey is organized as follows. Section 2 introduces the background and key concepts of multimodal video understanding. Section 3 organizes these concepts hierarchically according to the progressive structure of core capabilities in multimodal video understanding and summarizes representative methods for different tasks. Section 4 presents mainstream datasets, evaluation metrics, and representative benchmark results. Section 5 discusses the limitations of current approaches, application scenarios, and future research directions. Section 6 concludes the survey.

2. Background

2.1. Evolution of Video Understanding

The evolution of video understanding has progressed from low-level handcrafted features to deep spatio-temporal representation learning and, more recently, multimodal foundation models. Early studies primarily relied on manually designed descriptors to capture motion and appearance patterns. With the development of deep learning, CNN-based two-stream networks and 3D convolutional architectures became dominant approaches for learning visual video representations. Subsequently, self-attention architectures, Vision Transformer variants, and video-language pretraining paradigms further shifted the field from purely visual recognition toward scalable multimodal modeling and stronger semantic generalization. Figure 2 illustrates this developmental trajectory.

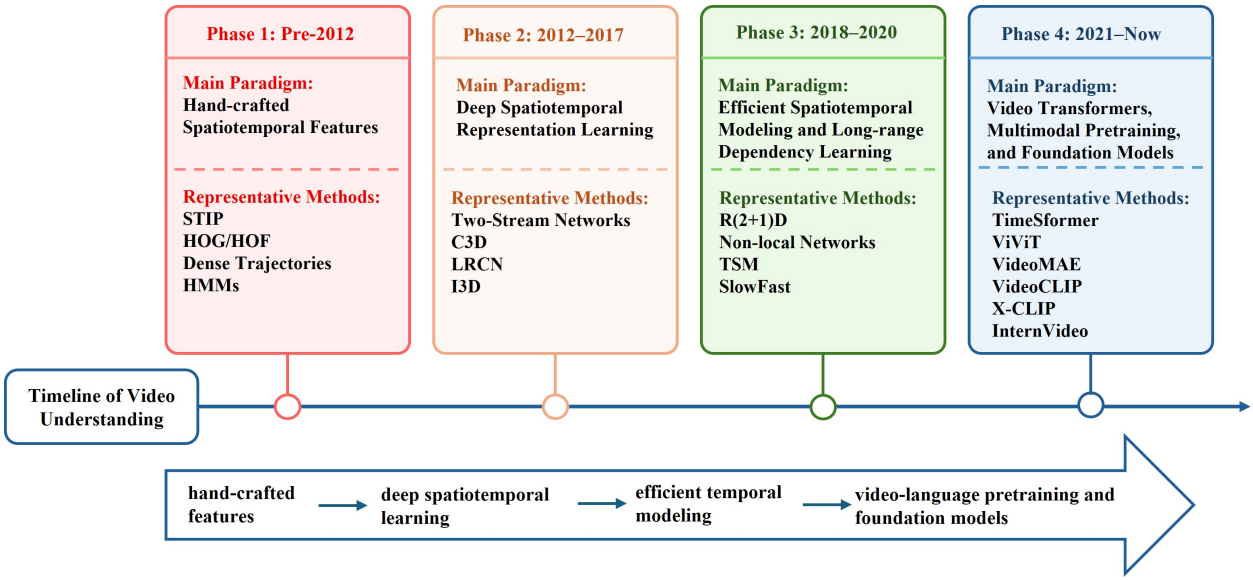

Figure 2. Historical evolution of video understanding. The timeline summarizes representative stages from handcrafted features to deep spatio-temporal learning, efficient temporal modeling, and video-language foundation models, with representative methods listed for each stage.

Early video understanding primarily relied on handcrafted spatio-temporal representations and shallow sequence models. Representative methods include spatio-temporal interest points (STIP) [14], Cuboids [15], dense trajectories [16], and hidden Markov models for temporal dependency modeling [17]. Although these methods are computationally efficient and interpretable, their representational capacity is limited, making them difficult to scale to large and diverse video datasets.

With the rise of deep learning, convolutional networks became the dominant paradigm for visual video understanding. Karpathy et al. demonstrated large-scale video classification using CNNs on Sports-1M [1], while two-stream networks modeled RGB appearance and optical-flow-based motion through separate streams [2]. Later, C3D [6] and I3D [18] further enhanced spatio-temporal modeling through 3D convolutions. These methods established the foundation for learning visual video representations and provided important encoders for subsequent multimodal systems.

Recent studies have increasingly adopted self-attention architectures and large-scale pretraining to improve global dependency modeling and generalization across datasets. Representative models include TimeSformer [19], ViViT [20], Video Swin [21], and VideoMAE [22]. In multimodal video understanding, these models often serve as visual backbones or video token generators. However, purely visual modeling still struggles to capture speech semantics, background sounds, and on-screen text, highlighting the need for multimodal joint modeling.

2.2. Modalities and Representations

Multimodal video understanding typically involves jointly modeling visual, audio, and textual information within a unified learning or reasoning system. These modalities provide complementary cues at different semantic levels: visual signals describe objects, scenes, actions, and interactions; audio signals provide cues related to speech, music, and environmental sounds; and textual signals offer explicit semantic information from ASR transcripts, subtitles, captions, or OCR-recognized text.

The visual modality is the primary source of information for video understanding. It captures both spatial cues, such as object appearance and scene layout, and temporal dynamics, such as motion patterns and human-object interactions. Visual representation learning has evolved from 2D CNNs for frame-level spatial encoding [23] to 3D CNNs for joint spatio-temporal modeling [18] and, more recently, to Vision Transformer-based architectures that capture long-range dependencies through self-attention [20].

Audio provides important non-visual information, especially when visual content is blurred, occluded, or semantically ambiguous. Speech can reveal dialogue content and communicative intent, while music and environmental sounds can indicate emotional tone, scene context, or sudden events. In representation learning, audio waveforms are often converted into spectrograms and processed by CNN- or Transformer-based encoders, such as VGGish [24], PANNs [25], and AST [26]. For speech-related information, self-supervised models such as Wav2Vec 2.0 are also widely used to extract semantic representations [27].

Textual information usually comes from ASR transcripts, manually captioned content, subtitles, or OCR-recognized screen text. Compared with visual and audio signals, text provides more explicit and abstract semantic descriptions, making it important for dialogue analysis, plot understanding, and entity reference. Text representation learning has shifted from static word embeddings and recurrent networks to Transformer-based language models. BERT and its variants learn contextual representations through bidirectional self-attention [28], while CLIP-like text encoders are commonly used to align textual semantics with visual representations in a shared embedding space [29].

2.3. Fusion and Cross-Modal Interaction

Multimodal fusion is a central component of multimodal video understanding. It aims to integrate and coordinate heterogeneous visual, audio, and textual information for more comprehensive semantic analysis. Compared with unimodal modeling, fusion enables models to exploit complementary information across modalities and

improve robustness under noisy or ambiguous conditions. However, because different modalities vary in data distribution, temporal granularity, and semantic abstraction, effective fusion requires not only feature integration but also cross-modal alignment and interaction modeling.

Existing fusion strategies are commonly categorized into early, late, and hybrid fusion. Early fusion integrates modality-specific features before prediction, typically through concatenation, summation, or weighted aggregation. This strategy is simple and can capture low-level correlations, but it often requires strict temporal alignment and may suffer from high-dimensional representations or modality dominance [30,31]. Late fusion combines the predictions of independently trained modality-specific models through averaging, voting, or a learnable decision module. This strategy is flexible and robust to missing modalities, but it limits deep cross-modal interaction during representation learning [30,31]. Hybrid fusion introduces modality interaction within intermediate network layers, often through lateral connections, tensor fusion, or cross-attention mechanisms. It has become a prevalent paradigm in modern Transformer-based multimodal systems because it preserves modality-specific information while enabling fine-grained semantic interaction [7,32]. Figure 3 summarizes these three fusion strategies.

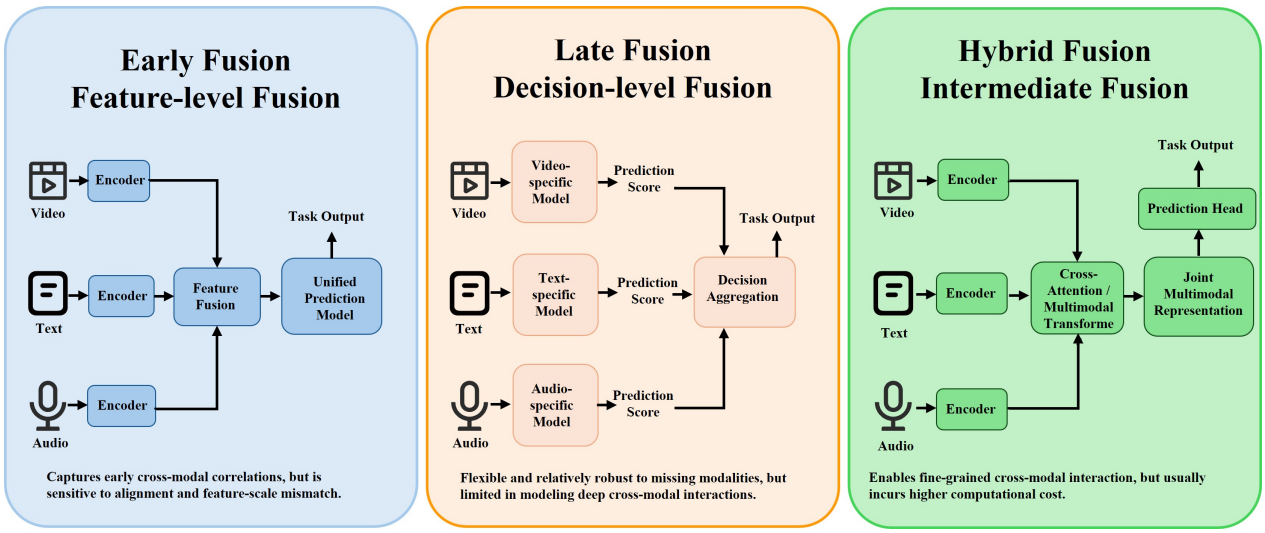

Figure 3. Comparison of fusion strategies for multimodal video understanding. Early fusion integrates modality features before prediction, late fusion combines modality-specific decisions, and hybrid fusion supports interaction within intermediate representation layers.

Despite advances in fusion mechanisms, cross-modal interaction remains challenging. First, modality heterogeneity makes it difficult to directly compare visual frames, acoustic signals, and discrete text tokens within a shared feature space. Second, semantic alignment is non-trivial because different modalities describe the same event with different information densities and levels of abstraction. For example, text may provide explicit semantic descriptions but omit visual details, while audio may contain contextual or event-related cues that are not strictly synchronized with visual content [33]. Third, modality missingness frequently occurs in practical scenarios, such as missing audio tracks, unavailable ASR transcripts, or failed OCR results. These issues challenge the robustness and generalization ability of multimodal video understanding systems.

These fusion-related issues also motivate a capability-oriented view of multimodal video understanding. Rather than treating multimodal modeling merely as a feature combination, it is necessary to examine how models align information across modalities, express video semantics in structured or generative forms, and support evidence-based

reasoning over complex video content. This consideration motivates the hierarchical framework adopted in the following sections.

3. Multimodal Methods for Video Understanding

The introduction of multimodal approaches has significantly improved video understanding by enhancing performance on existing tasks and enabling new forms of video comprehension. The core capabilities of multimodal video understanding can be summarized as three capability layers: first, establishing computable semantic correspondences between heterogeneous modalities; second, organizing video evidence into readable semantic representations; and finally, performing problem-oriented reasoning over multiple pieces of evidence and temporal dependencies. The theoretical justification for this framework is that these layers correspond to different computational roles in processing multimodal evidence, rather than to superficial differences in task names or benchmark formats. This hierarchy describes capability dependencies rather than chronological evolution. Therefore, this survey categorizes multimodal video understanding methods into three capability layers: alignment, expression, and reasoning. The proposed framework is not intended as a simple renaming of existing task categories such as video-text retrieval, video captioning, and video question answering. These conventional categories are mainly task-based: they organize studies according to benchmark settings and output formats. In contrast, the present framework is capability-based: it organizes methods according to the underlying computational role required for multimodal video understanding. Alignment refers to the establishment of computable and temporally grounded correspondences across heterogeneous modalities; expression refers to the selection and externalization of video evidence into faithful semantic descriptions; and reasoning refers to the integration of distributed multimodal evidence to support query-specific judgments or conclusions. From this perspective, retrieval, captioning, and VideoQA are representative task instances rather than the conceptual basis of the taxonomy. This distinction is necessary because modern MVU systems, especially large multimodal models, often blur task boundaries: captioning may require grounding, VideoQA may rely on retrieval or generated descriptions, and dense video understanding may involve alignment, expression, and reasoning simultaneously. A capability-based view therefore enables cross-task analysis of shared bottlenecks, such as weak temporal grounding, modality dominance, hallucination, missing modalities, and unsupported language priors, which are difficult to capture through isolated task buckets. It should be noted that this hierarchy does not imply a fixed pipeline of methods. This survey categorizes approaches based on primary research objectives, main output formats, and primary evaluation protocols. Methods that span multiple capability categories are assigned based on their primary contributions and main experimental settings. Figure 4 illustrates these three capability layers and their representative subtasks.

Although the three categories in Figure 4 are separated for clarity, they are not intended to be mutually exclusive. Some tasks involve multiple capabilities. For example, dense video captioning includes temporal localization, but it is categorized as expression because its final output is a set of event descriptions. Similarly, video question answering may rely on alignment or generated descriptions, but it is categorized as reasoning because its goal is to produce a question-specific answer or judgment. Thus, ambiguous tasks are assigned according to their dominant capability rather than all intermediate operations involved.

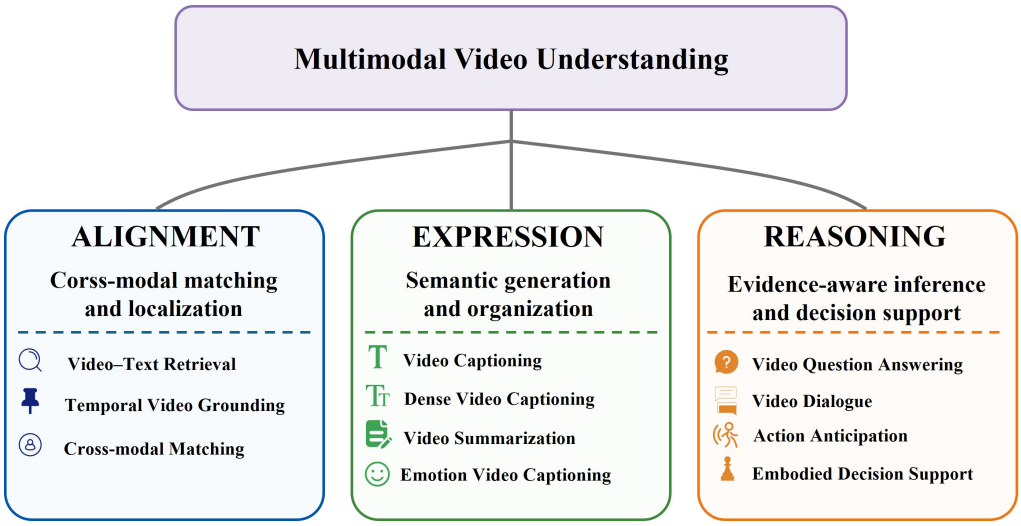

Figure 4. Taxonomy of multimodal video understanding tasks. The taxonomy groups representative tasks into alignment, expression, and reasoning layers according to their primary objectives, output formats, and evaluation protocols.

3.1. Alignment

The alignment layer focuses on establishing stable and computable semantic correspondences across heterogeneous modalities, such as visual, textual, and auditory signals. Its central role is to map information from different modalities into comparable representation spaces or explicit correspondence structures, thereby supporting matching, retrieval, grounding, and synchronization. In this layer, the primary objective is not to generate readable descriptions or infer final answers, but to determine whether two pieces of multimodal evidence are semantically consistent and where such consistency occurs. In other words, alignment-oriented methods address the following question: given a query in one modality, which content in another modality is most relevant to it, and where can this correspondence be localized? Accordingly, this survey regards a method as belonging to the alignment layer when its primary task centers on cross-modal matching, ranking, or alignment; its main outputs are structured results, such as similarity scores, ranked lists, matching matrices, or temporal grounding boundaries; and its evaluation mainly follows retrieval or alignment protocols, including Recall@K, average precision, or overlap-based localization metrics. Representative tasks include text-to-video retrieval, video-to-text matching, text-based temporal segment localization, and cross-modal synchronization.

The main difficulty in alignment is not merely combining video and text features, but learning a representation that is simultaneously transferable, temporally sensitive, and robust to noisy real-world supervision. Early Transformer-based retrieval studies therefore moved beyond feature concatenation toward contrastive and end-to-end representation learning. VATT and VideoCLIP showed that large-scale contrastive objectives could build shared video-language spaces from tri-modal or loosely paired data [34,35], while Frozen-in-Time improved scalability by unifying image and video encoders within an end-to-end training framework [36]. Once this shared-space paradigm became dominant, the technical focus shifted to the granularity of matching: T2VLAD introduced global-local retrieval, and GPO improved sequence-level aggregation for more stable semantic matching [37,38]. Studies such as SEA and MDMMT further indicate that alignment models must also handle text-side variation and domain shift, rather than assuming a clean correspondence between captions and video clips [39,40].

The emergence of CLIP changed the alignment problem from training a video-text model from scratch to adapting strong image-text priors to temporal video semantics. CLIP itself established a transferable dual-tower vision-language paradigm through large-

scale contrastive learning [29]; CLIP4Clip then demonstrated that this prior could be reused for video retrieval through lightweight temporal aggregation and fine-tuning [41]. However, direct transfer from images to videos leaves two unresolved gaps. The first is temporal degradation: methods such as X-CLIP, X-Pool, and VIOLET attempt to recover temporal or fine-grained correspondence through multi-granularity comparison, text-conditioned pooling, or masked visual-token modeling [42–44]. The second is supervision noise and system scalability: robust learning for partially mismatched pairs reduces the damage caused by noisy video-text pairs, while deep semantic hashing compresses retrieval representations for large-scale search [45,46]. These studies show that CLIP-style transfer is effective, but it does not remove the need for video-specific temporal modeling and deployment-aware retrieval design.

Recent large-model-based studies further extend alignment in two directions. One direction increases the capacity and openness of the representation foundation: InternVL scales vision-language representations and aligns them with large language models, while Open-VCLIP++ uses language-model-generated descriptions to strengthen open-vocabulary video retrieval [47,48]. The other direction models ambiguity more explicitly. Geometric Matching treats cross-modal semantics as regions rather than single embedding vectors, HCMI builds hierarchical interactions from frame-word relations to video-sentence relations, and GLSCL uses global-local semantic completion to reinforce both contextual and detailed alignment [49–51]. Therefore, the alignment layer has evolved from learning a common embedding space toward modeling fine-grained, uncertain, and scalable correspondence structures.

Alignment-oriented methods have effectively solved the problem of making heterogeneous video and language signals comparable for retrieval, matching, and grounding. Dual-encoder and CLIP-based models are especially useful in large-scale applications because they support offline indexing and efficient similarity search, while cross-attention and hierarchical interaction models provide finer semantic correspondence. However, high retrieval performance often reflects global semantic similarity rather than verified temporal evidence. These methods also remain sensitive to noisy captions, weak video-text pairing, domain shift, and missing modalities. Future research should therefore prioritize temporally grounded and uncertainty-aware alignment, enabling models to distinguish general relevance from evidence-level correspondence while balancing accuracy, robustness, and computational cost.

3.2. Expression

The expression layer focuses on transforming multimodal video evidence into readable and communicable semantic representations. Positioned between alignment and reasoning, this layer does not merely measure cross-modal similarity, nor does it primarily aim to answer a specific query. Instead, it requires models to select salient visual-temporal evidence, organize events or segments, and externalize video content through natural language or structured semantic descriptions. Compared with alignment-oriented methods, expression-oriented methods place greater emphasis on what should be described, how event boundaries should be organized, and whether the generated representation remains temporally coherent. This survey considers a method to belong to the expression layer when its primary task centers on generative semantic externalization; its main outputs are captions, event descriptions, summaries, or related textual representations; and its evaluation focuses on generation quality, semantic consistency, event coverage, redundancy control, or temporal coherence. Representative tasks include video captioning, dense video captioning, emotional or domain-specific captioning, and semantic summarization from video to text.

The central tension in expression-oriented methods is between end-to-end generation and controllable event organization. End-to-end models reduce the dependence on hand-designed detection or proposal pipelines, but they may also lose fine-grained temporal details or produce descriptions with weak event structure. PDVC established an end-to-end dense captioning framework, and SwinBERT showed that pure Transformer architectures could directly model sampled video patches densely while controlling sequence cost through sparse attention [52,53]. Subsequent work did not simply abandon structure; instead, it reintroduced structure where generation needed stronger temporal coherence. Hierarchical scene-event-frame or video-shot-frame designs improve long-range organization, linguistic modularization and long-short-term relationship graphs strengthen relational modeling, and dynamic semantic grouping reduces redundant inter-frame information [54–58]. Efficiency and coverage form another constraint: non-autoregressive generation and compressed domain features accelerate inference, while external knowledge graphs help mitigate limitations of long-tail vocabulary [59–61]. Thus, expression models develop through a recurring balance between architectural simplification and explicit semantic control.

Generative pretraining and multimodal large models shift the expression layer from task-specific captioning toward unified sequence generation and long-video semantic externalization. Vid2Seq treats dense event localization and description as a specialized sequence-generation problem, demonstrating that large-scale weak supervision can partially shift the localization burden onto autoregressive modeling [62]. For longer videos, however, the main bottleneck is memory rather than caption quality alone. MA-LMM introduces online memory to extend existing multimodal LLMs without linearly increasing context length, while Streaming-DVC combines streaming decoding with clustering memory, enabling event descriptions to be produced incrementally rather than only after offline processing [63,64]. Image-Text Transfer further suggests that pretrained image foundation models can support video description with lightweight adaptation or frame concatenation [65]. These methods collectively indicate that expression is moving from sentence-level description to temporally sustained semantic transformation under computational and context constraints.

At the same time, large general-purpose models do not eliminate the need for task-specific guidance. Emotion-aware and vertical-domain captioning require the model to select subtle cues that are often underrepresented in generic pretraining data. ECPA uses prompt learning to guide frozen models toward facial microexpressions and emotional semantics, while emotion-state guidance improves the coverage and consistency of affective descriptions [66,67]. In domain-specific scenarios, EIKA combines explicit player knowledge with visual scene knowledge for sports description, and ST-GIN models spatio-temporal teacher-student interactions in classroom videos [68,69]. Memory-based modeling and image-composite enhancement further improve semantic richness and robustness in complex scenes [70,71]. Therefore, the expression layer should be understood not as generic caption generation, but as grounded semantic externalization: models must decide what evidence is worth expressing, how events should be organized, and how domain knowledge should constrain the generated text.

Expression-oriented methods have advanced video understanding from recognition and retrieval to human-readable semantic generation. They are useful for captioning, dense video description, summarization, and domain-specific reporting. End-to-end captioning models simplify short-video description, whereas dense, hierarchical, memory-based, and streaming methods are more suitable for long or complex videos. Nevertheless, expression methods still face a major tension between fluency and grounding. Common metrics such as BLEU, METEOR, ROUGE-L, and CIDEr mainly measure textual overlap, but

they do not fully evaluate factual correctness, temporal coherence, or event coverage. Future work should focus on evidence-grounded expression, where models identify salient events, generate temporally consistent descriptions, and reduce hallucinated or weakly supported content.

3.3. Reasoning

The reasoning layer focuses on producing query-specific answers or evidence-supported judgments by integrating temporal, relational, and multimodal cues. Unlike the expression layer, which emphasizes the generation of descriptive passages, the reasoning layer is primarily concerned with whether a model can identify question-relevant evidence, integrate information across temporal segments or entity relationships, and derive a correct conclusion. This process often requires multi-step evidence aggregation rather than direct recognition of visible objects or actions. Accordingly, this survey regards a method as belonging to the reasoning layer when its primary task centers on question answering or conclusion-oriented judgment; its main output is an answer, decision, or evidence-supported response rather than a ranked list or open-ended narrative description; and its evaluation mainly focuses on answer correctness, reasoning consistency, and, when applicable, the reliability or interpretability of evidence localization. Representative tasks include video question answering, temporal reasoning, causal reasoning, multi-hop reasoning, and evidence-based long-form video question answering.

Reasoning-oriented methods first need to solve an evidence organization problem: before a model can answer a question, it must identify which objects, actions, temporal segments, and modality cues are relevant. Early studies therefore emphasized explicit spatio-temporal dependency modeling. Bridge to Answer represents appearances and actions as heterogeneous graph nodes, CoVGT uses a Video Graph Transformer to capture dynamic object interactions, and TASTA applies text-assisted spatio-temporal attention to suppress irrelevant background information [72–74]. These approaches differ architecturally, but they share the assumption that reasoning quality depends on making video evidence more query-aware and relationally organized. Later work extends this idea by separating static and dynamic semantics or by selecting evidence iteratively in long videos. Semantic decoupling aligns noun-related static frames with verb-related motion features, while MIST reduces long-video information overload by iteratively selecting key segments [75,76]. This line of research shows that VideoQA is not only a language-answering problem but also a problem of evidence filtering and temporal abstraction.

Another line of reasoning questions whether high benchmark accuracy actually reflects causal understanding. NExT-QA exposed the weakness of existing models in causal action reasoning and pushed evaluation from describing visible content toward explaining why events occur [77]. Subsequent causal and counterfactual methods target spurious dataset correlations more directly. IGV uses causal intervention to separate critical scenarios from environmental bias, causal-diagram-based modeling decouples visual confounders from textual keyword bias, and cross-modal contrastive learning constructs negative pairs to reduce linguistic shortcut learning [78–80]. The same principle also extends beyond standard VideoQA: counterfactual reasoning in multimodal speech anxiety detection suggests that eliminating spurious correlations can improve the reliability of multimodal judgment in domain-specific settings [81]. Compared with alignment and expression, the reasoning layer therefore places stronger emphasis on robustness, interpretability, and whether the model's answer follows from valid evidence rather than dataset regularities.

Large language models have recently changed the implementation form of video reasoning, but they have not removed the core need for evidence control. CLIPBERT showed that sparse visual sampling and end-to-end visual-language learning could support effi-

cient question answering, MERLOT introduced self-supervised event-order learning to acquire script-like knowledge, and Frozen BiLM demonstrated the value of frozen language models for zero-shot VideoQA [82–84]. Data scaling through the Just Ask paradigm further improves generalization by generating large-scale QA supervision from narrated videos [85]. In the LLM era, the dominant challenges become long-context management and multi-step controllability: LongVLM compresses long videos through hierarchical token merging, MoReVQA uses an LLM as a controller for modular tool-based reasoning, and prompt-based strategies guide frozen models toward fine-grained clues without full fine-tuning [86–88]. Nevertheless, empirical analysis of Video-LLMs shows that fluent generation can coexist with weak temporal reasoning and limited adversarial robustness [89]. Thus, future reasoning systems must combine the flexibility of large models with explicit evidence grounding, temporal verification, and hallucination control.

Reasoning-oriented methods address the need to produce query-specific answers or judgments from temporally distributed multimodal evidence. Graph-based, attention-based, causal, counterfactual, modular, and large-model-based approaches have improved evidence organization and flexible question answering. However, high VideoQA accuracy does not necessarily indicate genuine temporal or causal reasoning, because models may exploit language priors, answer biases, short-term visual cues, or dataset shortcuts. Large multimodal models may also produce fluent but weakly grounded answers. Future research should prioritize evidence-controlled reasoning, enabling models to locate supporting visual, audio, or textual evidence, distinguish causal inference from correlation, and provide reliable uncertainty estimates when evidence is insufficient.

4. Datasets and Benchmarks

4.1. Datasets

Table 3 summarizes representative datasets used in multimodal video understanding, covering alignment, expression, and reasoning-oriented tasks.

Table 3. Summary of representative datasets for multimodal video understanding.

Dataset	Mod.	Scale	Typical Tasks	Key Characteristics/Usage Notes
MSR-VTT [90]	V + T	10 K clips; about 200 K sentences	Retrieval; captioning	Open-domain benchmark with standard captioning and 1 K retrieval settings.
VATEX [91]	V + T	41.3 K clips; about 826 K captions	Multilingual captioning; MT	Bilingual English/Chinese captions over 600 action categories.
HowTo100 M [92]	V + T	1.22 M videos; about 136 M pairs	Pretraining; retrieval transfer	Large-scale ASR supervision with noisy temporal pairing.
MSVD [93]	V + T	1970 videos; about 80 K sentences	Video captioning	Classic short-video captioning benchmark; common 1200/100/670 split.
ActivityNet Captions [94]	V + T	About 20 K videos; 100 K event captions	Dense captioning; grounding	Long-video benchmark with temporally grounded event descriptions.
CMU-MOSEI [95]	V + A + T	23,453 utterance clips	Sentiment; emotion analysis	Aligned text, visual, and audio signals with affective labels.
MELD [96]	V + A + T	1433 dialogues; about 13 K utterances	Dialogue emotion recognition	Multi-speaker dialogue benchmark with emotion and sentiment labels.
TGIF-QA [97]	V + T	About 72 K GIFs; 165 K QA pairs	VideoQA; counting	Short dynamic clips for repetition, state transition, and FrameQA.
NExT-QA [77]	V + T	5440 videos; about 52 K QA pairs	VideoQA; causal reasoning	Focuses on causal, temporal, and descriptive questions.

Note: V, A, and T denote vision, audio, and text, respectively. Sizes and splits follow commonly reported settings and may vary across preprocessing or evaluation protocols.

MSR-VTT, VATEX, MSVD, and ActivityNet Captions are widely used for video-text retrieval, captioning, dense captioning, and temporal grounding [90,91,93,94]. MSR-VTT and MSVD mainly support short-video retrieval and captioning, VATEX extends video-language evaluation to bilingual captioning and translation-oriented settings, and ActivityNet Captions emphasizes temporally grounded event description in longer videos. These

datasets provide useful benchmark anchors, but they differ substantially in annotation density, temporal granularity, language source, and evaluation split, which limits direct cross-dataset comparison.

For large-scale pretraining, HowTo100M provides weakly aligned video-text supervision from narrated instructional videos [92]. Its scale and diversity make it valuable for representation learning, but its ASR-based supervision also introduces noisy temporal pairing and incomplete language descriptions. For affective and dialogue-oriented understanding, CMU-MOSEI and MELD provide tri-modal emotion or sentiment annotations [95,96]. These datasets are useful for studying audio-visual-text interaction, but their results should be interpreted with attention to annotation subjectivity, class imbalance, speaker or domain bias, and context segmentation.

For reasoning-oriented evaluation, TGIF-QA and NExT-QA focus on dynamic, temporal, and causal question answering [77,97]. TGIF-QA emphasizes short dynamic clips and repetitive or state-transition questions, whereas NExT-QA shifts evaluation toward causal and temporal reasoning. However, both remain bounded benchmark settings, and performance on them does not necessarily imply robust long-video reasoning in open-world scenarios.

4.2. Metrics and Performance Comparison

The benchmark results summarized in this section are collected from the original publications and should be interpreted as reported results rather than as a unified leaderboard. Even on the same dataset, results may be affected by differences in data splits, pretraining corpora, modality inputs, backbone architectures, fine-tuning strategies, and evaluation protocols. Therefore, comparisons are most meaningful within the same task, dataset setting, and evaluation protocol.

4.2.1. Benchmark Analysis of Alignment Methods

In alignment-based tasks, the model's core objective is to measure semantic consistency between videos and text and to perform matching, ranking, or alignment based on this consistency. Consequently, its evaluation metrics primarily focus on retrieval ranking quality and discrimination capability in matching, rather than generation quality or inference accuracy. The most commonly used metric is Recall@K, which measures whether correct matches appear among the top-K candidates returned by the model for a given query. For example, let the query text set be Q. For each text $q \in Q$, the model ranks all candidate videos by similarity. If the corresponding true video appears within the top-K positions, it is counted as a hit [10]. Recall@K is defined as:

$$\text{R@}K = \frac{1}{|Q|} \sum_{q \in Q} \mathbb{I}(\text{rank}(v_q) \leq K) \tag{1}$$

Here, $\mathbb{I}(\cdot)$ denotes the indicator function, and $\text{rank}(v_q)$ represents the position of the ground-truth video corresponding to query q in the ranked results. The computation of video-to-text retrieval is fully symmetric, with only the query and candidate sets swapped. R@1, R@5, and R@10 are the most commonly reported configurations, with R@1 being the most sensitive to model discrimination capability.

In addition to Recall@K, Median Rank (MdR) is often used as a supplementary metric to characterize the overall stability of a model's ranking performance [10]. MdR is defined as the median rank of all ground-truth matches:

$$\text{MdR} = \text{median}\{\text{rank}(v_q) \mid q \in Q\} \tag{2}$$

Compared with Recall@K, MdR is more sensitive to ranking errors in the long tail; a smaller value indicates better overall retrieval quality.

Table 4 reports representative alignment results.

Table 4. Reported performance of representative alignment methods.

Year	Method	Evaluation Setting	R@1	R@5	R@10	MdR ↓
MSR-VTT (1k-A)						
2021	VATT [34]	MSR-VTT (1k-A)	–	–	23.6	67
	Frozen in Time [36]	MSR-VTT (1k-A)	31.0	59.5	70.5	3
	MDMMT [40]	MSR-VTT (1k-A)	38.9	69.0	79.7	2
	SEA [39]	MSR-VTT (1k-A)	23.8	50.3	63.8	5
	T2VLAD [37]	MSR-VTT (1k-A)	29.5	59.0	70.1	4
	VideoCLIP [35]	MSR-VTT (1k-A)	30.9	55.4	66.8	–
2022	CLIP4Clip [41]	MSR-VTT (1k-A)	44.5	71.4	81.6	2
	VIOLET [44]	MSR-VTT (1k-A)	34.5	63.0	73.4	–
	X-Pool [43]	MSR-VTT (1k-A)	43.9	72.5	82.3	2
	X-CLIP [42]	MSR-VTT (1k-A)	49.3	75.8	84.8	2
2023	RCL [45]	MSR-VTT (1k-A)	17.7	44.2	57.2	7
2025	HCMI [50]	MSR-VTT (1k-A)	55.0	80.4	86.8	1
	Geometric Matching [49]	MSR-VTT (1k-A)	24.5	52.5	64.2	–
Other reported settings						
2021	VSE∞ (GPO) [38]	MSR-VTT 6573/2990/497 split	8.7	25.3	35.9	–
2025	GLSCL [51]	MSR-VTT, setting not specified	43.2	76.0	86.7	–

Note: The downward arrow (↓) indicates that lower values are preferred for MdR. Results are grouped by evaluation setting. The MSR-VTT (1k-A) block is the main relatively comparable subset, but even within this block, methods may differ in pretraining data, backbone architecture, CLIP or language-model initialization, retrieval direction, and fine-tuning settings. Results outside this setting are included only as references and should not be interpreted as strict head-to-head comparisons. VSE∞ (GPO) is listed outside the MSR-VTT (1k-A) block because the original paper reports it on the MSR-VTT standard split for text-to-video retrieval, rather than the 1k-A setting.

4.2.2. Benchmark Analysis of Expression Methods

In expression-based tasks, the model's objective is to generate natural-language descriptions that capture the primary semantic content of a video. Consequently, its evaluation metrics primarily measure the similarity between generated text and reference captions in terms of lexical overlap, semantic consistency, and content coverage. The evaluation paradigm has shifted from ranking-based metrics to generative quality assessment metrics.

BLEU was the earliest widely adopted generative evaluation metric, focusing on statistical measurement of exact n-gram matches between generated and reference captions. It reflects local word order and phrase consistency but is sensitive to synonym substitutions and semantically equivalent expressions, often undervaluing linguistic diversity [98]. BLEU is defined as:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \cdot \ln(p_n)\right) \tag{3}$$

p_n denotes n-gram precision, w_n represents the corresponding weight, and BP signifies the brevity penalty. This metric emphasizes local word order matching and is insensitive to semantically equivalent expressions.

METEOR explicitly balances precision and recall on top of BLEU and mitigates the issue of "literal inconsistency but semantic equivalence" through stemming and synonym matching. It is generally considered to correlate more strongly with human judgments [99]. METEOR is defined as:

$$\text{METEOR} = F_\alpha \cdot (1 - \text{Penalty}) \tag{4}$$

F_α represents the weighted F-score; this metric places greater emphasis on semantic consistency and typically correlates more closely with human evaluation.

ROUGE-L measures the consistency between generated captions and reference captions in overall syntactic structure based on the longest common subsequence. It is sensitive to sentence-order preservation and is often used to address the shortcomings of n-gram metrics in modeling long-sentence structures [100]. ROUGE-L is defined as:

$$R_{\text{lcs}} = \frac{\text{LCS}(C, R)}{|R|} \tag{5}$$

$$P_{\text{lcs}} = \frac{\text{LCS}(C, R)}{|C|} \tag{6}$$

$$\text{ROUGE-L} = F_{\text{lcs}} = \frac{(1+\beta^2) R_{\text{lcs}} P_{\text{lcs}}}{R_{\text{lcs}} + \beta^2 P_{\text{lcs}}} \tag{7}$$

where C and R denote the generated caption and the reference caption, respectively. $\text{LCS}(C, R)$ denotes the length of their longest common subsequence. R_{lcs} and P_{lcs} represent the LCS-based recall and precision, and β controls the relative weight between recall and precision.

CIDEr represents captions as TF-IDF-weighted n-gram vectors and emphasizes "human consensus descriptions" through multi-reference consistency modeling. It is regarded as one of the most representative metrics for content coverage and semantic adequacy in video caption generation and has become a core reporting metric on multiple mainstream benchmarks [101]. CIDEr is defined as:

$$\text{CIDEr} = \frac{1}{M} \cdot \sum_{m=1}^{M} \frac{\boldsymbol{g} \cdot \boldsymbol{r}_m}{\|\boldsymbol{g}\| \cdot \|\boldsymbol{r}_m\|} \tag{8}$$

Here, $\boldsymbol{g}$ represents the generated caption vector, and $\boldsymbol{r}_m$ denotes the mth reference caption. This metric emphasizes content coverage aligned with human consensus, serving as a primary indicator in video caption generation.

In addition to traditional lexical-overlap metrics, recent evaluation protocols have increasingly considered learned semantic and vision-language metrics, such as BERTScore and CLIPScore [102,103]. BERTScore evaluates the semantic similarity between a generated caption and reference captions using contextual token embeddings. Given a candidate sentence $C = \{\boldsymbol{c}_i\}_{i=1}^{m}$ and a reference sentence $R = \{\boldsymbol{r}_j\}_{j=1}^{n}$, their token embeddings are first extracted from a pretrained language model. The BERTScore precision, recall, and F1 score can be written as:

$$P_{\text{BERT}} = \frac{1}{m} \sum_{i=1}^{m} \max_j \cos(\boldsymbol{c}_i, \boldsymbol{r}_j), \tag{9}$$

$$R_{\text{BERT}} = \frac{1}{n} \sum_{j=1}^{n} \max_i \cos(\boldsymbol{r}_j, \boldsymbol{c}_i), \tag{10}$$

$$F_{\text{BERT}} = \frac{2 P_{\text{BERT}} R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}. \tag{11}$$

Compared with n-gram overlap metrics, BERTScore is less sensitive to surface-form variation and can better capture semantic equivalence between different expressions. However, it is still reference-dependent and does not directly verify whether the generated caption is grounded in the visual content.

CLIPScore evaluates cross-modal compatibility by measuring the similarity between visual and textual embeddings extracted from a pretrained vision-language model. Given a visual input V and a generated caption C, CLIPScore can be defined as:

$$\text{CLIPScore}(V, C) = w \cdot \max(\cos(E_V(V), E_T(C)), 0), \tag{12}$$

where $E_V(\cdot)$ and $E_T(\cdot)$ denote the visual and textual encoders of CLIP, respectively, and w is a scaling factor. Unlike reference-based metrics, CLIPScore can be used as a reference-free metric when human captions are limited or unavailable. Nevertheless, it may inherit biases from the pretrained vision-language encoder and may be insensitive to fine-grained temporal errors in videos. Therefore, future MVU evaluation should combine lexical-overlap metrics, learned semantic metrics, cross-modal grounding metrics, and human or evidence-based assessment.

Table 5 summarizes representative expression-oriented methods. Since VC, DVC, and EVC correspond to different generation objectives, the results are organized by task type and dataset setting.

Table 5. Reported results of representative expression-oriented methods.

Year	Method	Task	Dataset Setting	BLEU-4	METEOR	ROUGE-L	CIDEr
Video Captioning (VC) on MSVD							
2021	O2NA [59]	VC	MSVD	55.4	37.4	74.5	96.4
	SGN [58]	VC	MSVD	52.8	35.5	72.9	94.3
2022	LSRT [57]	VC	MSVD	55.6	37.1	73.5	98.5
	SwinBERT [53]	VC	MSVD	66.3	42.4	80.9	149.4
	HMN [56]	VC	MSVD	59.2	37.7	75.1	104.0
2023	Compressed [60]	VC	MSVD	60.1	41.4	78.2	121.5
2024	MAN [70]	VC	MSVD	59.7	37.3	74.3	101.5
	GLG [104]	VC	MSVD	63.6	42.4	80.8	107.8
Video Captioning (VC) on MSR-VTT							
2024	IcoCap [71]	VC	MSR-VTT	47.0	31.1	64.9	60.2
2025	PIT [65]	VC	MSR-VTT	52.4	34.2	68.3	79.5
Video Captioning (VC) on other datasets							
2023	TextKG [61]	VC	YouCookII	11.7	18.4	40.2	133.0
2024	MA-LMM [63]	VC	YouCookII	–	17.6	–	131.2
2025	EIKA [68]	VC	VC-NBA-2022	36.7	29.5	56.8	140.7
Dense Video Captioning (DVC)							
2021	PDVC [52]	DVC	ActivityNet	3.12	11.26	–	53.65
	ECHR [54]	DVC	ActivityNet	1.96	10.58	–	39.73
2023	Vid2Seq [62]	DVC	YouCookII	–	9.3	–	47.1
2024	Streaming DVC [64]	DVC	ActivityNet	–	9.0	–	41.2
Emotional Video Captioning (EVC) on EmVidCap							
2024	VEIN [67]	EVC	EmVidCap	27.1	21.6	46.8	39.4
2025	ECPA [66]	EVC	EmVidCap	21.73	21.6	38.53	20.33

Note: VC, DVC, and EVC denote video captioning, dense video captioning, and emotional video captioning, respectively. Comparisons should be made only within the same task, dataset, split, and evaluation protocol. Captioning scores are sensitive to tokenization, reference captions, decoding strategy, localization protocol for DVC, and evaluation scripts; therefore, the table summarizes reported trends rather than providing a unified ranking.

4.2.3. Benchmark Analysis of Reasoning Methods

In reasoning tasks, the model aims to generate verifiable answers to given questions using spatio-temporal evidence from videos. Consequently, VideoQA evaluation metrics center on answer correctness, leading to two mainstream paradigms that differ in answer space and annotation formats. First, discrete discriminative evaluation within a closed answer space primarily measures consistency between predicted and ground-truth answers. Second, consensus evaluation of open-ended textual answers accounts for natural-language diversity by assigning soft scores to plausible answers based on multi-annotator agreement. Unlike subtitle generation, which emphasizes lexical overlap and content coverage, VideoQA metrics prioritize task-oriented correctness and may include supplementary measures such as top-K tolerance and exact matching when required.

Table 6 summarizes representative reasoning-oriented VideoQA methods.

Table 6. Reported results of representative video question answering methods.

Year	Method	Dataset Setting	Accuracy (%)
MSR-VTT-QA			
2021	CLIPBERT [82]	MSR-VTT-QA	37.4
	MERLOT [83]	MSR-VTT-QA	43.1
2025	LongVLM [86]	MSR-VTT-QA	59.8
MSVD-QA			
2021	BTA [72]	MSVD-QA	37.2
	Just Ask (ICCV) [105]	MSVD-QA	46.3
2022	IGV [78]	MSVD-QA	40.8
	FrozenBiLM [84]	MSVD-QA	33.8
2025	Just Ask (TPAMI) [85]	MSVD-QA	47.5
TGIF-QA			
2023	TASTA [74]	TGIF-QA	58.1
NExT-QA			
2023	MIST [76]	NExT-QA	57.18
	CoVGT [73]	NExT-QA	60.73
	MCR [79]	NExT-QA	52.35
2024	DRV [75]	NExT-QA	57.05
	MoReVQA [87]	NExT-QA	69.2
	CCRL [80]	NExT-QA	51.8
2025	HeurVidQA [88]	NExT-QA	60.9

Note: Results are reported from the original papers for each dataset setting. Comparisons should be made only within the same dataset and protocol because accuracy depends on question type distribution, answer space, multiple-choice or open-ended format, data split, modality inputs, use of subtitles or ASR, external pretraining data, and large-model components.

Accuracy is the most commonly used primary metric in closed-domain VideoQA, measuring a model's discriminative correctness within a discrete answer space. This metric treats each question as a classification decision, counting predictions as correct only when they perfectly match the labeled answer and as incorrect otherwise. Its advantages include clear semantic interpretation and direct reflection of task success rates. However, it may be overly stringent when answer categories are highly imbalanced or when labels contain multiple acceptable expressions. Accuracy is defined as:

$$\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{a}_i = a_i) \tag{13}$$

N denotes the number of questions. $\hat{a}_i$ and a_i represent the predicted answer and the ground-truth answer for the i-th question, respectively. $\mathbb{I}(\cdot)$ is the indicator function.

5. Discussion

5.1. From Task-Specific Optimization to Capability-Oriented Understanding

Across alignment, expression, and reasoning tasks, multimodal video understanding has moved from task-specific neural architectures toward capability-oriented modeling. Earlier systems mainly optimized a single objective, such as retrieval ranking, caption generation, or answer classification. Their architectural designs therefore tended to emphasize one dominant component, such as cross-modal embedding learning in retrieval models, encoder-decoder generation in captioning models, or attention-based evidence aggregation in VideoQA models. Recent methods increasingly reuse pretrained image-text or video-language representations and then adapt them to video-specific temporal structure, as illustrated by CLIP-based retrieval models, Transformer-based captioning models, and large-model-based VideoQA systems [41,42,53,82,87].

This trend does not imply that the three capability layers have become identical. Alignment, expression, and reasoning still impose different neural computation requirements. Alignment requires discriminative representations that preserve fine-grained correspondence between visual, textual, and sometimes audio cues. Expression requires models to select events, organize temporal information, and generate faithful natural-language descriptions. Reasoning further requires evidence aggregation across clips and the suppression of dataset shortcuts or unsupported language priors. Therefore, the proposed capability-based view should be understood not only as a task taxonomy, but also as a way to separate the computational roles of representation learning, semantic generation, and evidence-based inference.

5.2. Cross-Modal Representation and Temporal Evidence Modeling

The central technical difficulty in multimodal video understanding is the joint modeling of heterogeneous modalities under temporal constraints. Visual frames, audio signals, subtitles, ASR transcripts, and OCR text differ in their sampling rates, semantic densities, and noise levels. Fusion strategies such as early, late, and hybrid fusion, as well as cross-attention, offer different trade-offs between modality-specific preservation and cross-modal interaction [7,30–32]. However, simple fusion is insufficient when the relevant evidence is sparse, temporally delayed, or distributed across multiple modalities.

Temporal evidence modeling is therefore a more specific bottleneck than generic multimodal fusion. Retrieval models must distinguish global semantic similarity from moment-level correspondence; captioning models must decide which events are salient enough to describe; and VideoQA models must identify evidence that supports a particular answer rather than relying on question priors. Long-video methods such as MA-LMM, Streaming-DVC, LongVLM, and MoReVQA address this issue through memory mechanisms, streaming or hierarchical processing, and modular reasoning strategies [63,64,86,87]. These approaches suggest that future progress will depend less on merely increasing the number of sampled frames and more on learning when to preserve, compress, retrieve, or discard temporal information.

5.3. Applications and Engineering Considerations

At the application level, the three capability layers of multimodal video understanding correspond to different engineering functions and deployment requirements. Alignment-oriented tasks, including text-to-video retrieval, video-to-text matching, temporal grounding, and cross-modal matching, mainly support systems that require fast access to large-

scale video content, such as short-video search engines, media asset management, surveillance review, and enterprise video archives. In such systems, video embeddings are often precomputed offline and stored in a vector database, while user queries are encoded online and matched through approximate nearest-neighbor search. Representative methods such as VideoCLIP, CLIP4Clip, and X-CLIP improve semantic matching by aligning visual and textual representations in a shared embedding space [35,41,42]. However, practical deployment should not be evaluated solely by retrieval accuracy. Indexing cost, query latency, storage overhead, update frequency, robustness to domain shift, and privacy-preserving handling of faces, speech transcripts, license plates, or OCR text are also central system-level concerns.

Expression- and reasoning-oriented tasks extend this engineering perspective from content access to semantic transformation and evidence-based assistance. Video captioning, dense video captioning, summarization, and emotional captioning can support subtitle generation, lecture summarization, sports highlight description, meeting analysis, and accessibility services, but they require different trade-offs between generation quality, temporal coverage, memory usage, and inference efficiency [53,62–64]. Reasoning-oriented tasks, including VideoQA, video dialogue, event explanation, and action anticipation, impose stricter requirements on reliability and interpretability because systems are expected to provide not only fluent answers but also traceable evidence from video content [76,86,87,89]. Therefore, deployable multimodal video systems should report system-level indicators such as P95 latency, throughput, GPU memory, indexing and storage costs, model update frequency, privacy-sensitive modality handling, confidence estimation, and failure-mode attribution. Such attribution is especially important when errors arise from missing modalities, ASR or OCR failures, corrupted audio, camera motion, domain shift, misleading content, or unsupported language priors.

Computational cost and sustainability should also be considered when deploying MVU systems. Recent models often improve performance by increasing parameter scale, using more video frames or tokens, introducing larger language decoders, or adding memory and tool-based modules. These designs may improve retrieval, captioning, or VideoQA accuracy, but they also increase GPU memory usage, inference latency, storage demand, and energy consumption. Therefore, practical evaluation should not compare models only by task performance, but should also report cost-related indicators such as parameter count, FLOPs or processed tokens, number of sampled frames, GPU memory, latency, throughput, and indexing or storage overhead. Since different studies may count parameters differently, for example by including or excluding frozen visual encoders, audio encoders, or large language models, these indicators should be reported with clear boundaries. Future benchmarks should encourage performance–cost analysis, such as accuracy or Recall@K under comparable computational budgets, to identify models that offer better trade-offs between effectiveness, efficiency, and sustainable deployment.

5.4. Algorithmic Bias and Ethical Implications in Emotion Interpretation

When MVU systems are applied to surveillance review, workplace monitoring, or corporate video archives, emotion interpretation should be treated with particular caution. Unlike retrieval or object recognition, emotion-related outputs infer latent affective states from facial expressions, speech prosody, body movement, dialogue content, and contextual cues, which are inherently ambiguous and may vary across individuals, cultures, languages, and recording conditions. Algorithmic bias may arise from imbalanced training data, simplified emotion annotations, acted or domain-specific datasets, ASR/OCR errors, facial occlusion, lighting conditions, accents, dialects, or the dominance of a single modality during fusion. In such scenarios, inferred emotion labels may affect security review,

employee evaluation, customer profiling, or risk assessment, even when the underlying evidence is uncertain. Therefore, emotion-related MVU outputs should be presented as uncertain and context-dependent interpretations rather than objective measurements of mental states, and should not be used as the sole basis for high-stakes decisions. Responsible deployment should include human review, uncertainty reporting, dataset and annotation documentation, robustness checks across relevant conditions, purpose limitation, data minimization, access control, retention limits, and auditability.

5.5. Evaluation Reliability and Benchmark Comparability

The reported progress of multimodal video understanding should be interpreted with caution because datasets, splits, metrics, and supervision settings are highly heterogeneous. Retrieval performance is usually reported with Recall@K and median rank, but these metrics do not directly measure whether the model localizes the correct temporal evidence. Captioning metrics such as BLEU, METEOR, ROUGE-L, and CIDEr quantify lexical or n-gram overlap to different degrees, but they may fail to fully capture factual grounding, event coverage, or temporal coherence. VideoQA accuracy is intuitive, but it can be affected by answer-space design, question type distribution, and language bias. These limitations are consistent with the metric analysis and the task-specific performance tables summarized in this survey.

Benchmark comparability is especially challenging for survey-level analysis. The same dataset name may correspond to different evaluation splits or task variants, and methods may differ in whether they use zero-shot transfer, task-specific fine-tuning, external pretraining data, or large language model components. Consequently, the performance tables in this survey should not be read as a single universal ranking across all methods. A more reliable evaluation practice is to report results within clearly specified dataset settings and to separate perception, alignment, generation, and reasoning errors whenever possible. For large multimodal models, this decomposition is particularly important because fluent language output can mask failures in temporal grounding or visual evidence selection [89].

For future evaluation of Video-LLMs, it is important to move beyond conventional text-overlap metrics and answer accuracy, because fluent responses may still be temporally misplaced or unsupported by video evidence. For temporal alignment, tIoU-based Recall@K and mean IoU, which are widely used in temporal sentence grounding and moment retrieval, can be adopted to evaluate whether the model grounds its output in the correct video segment [106]. For evidence fidelity and hallucination analysis, video-specific benchmarks such as VidHalluc provide a useful reference by evaluating hallucinations related to actions, temporal sequences, and scene transitions in multimodal large language models [107]. Therefore, future Video-LLM benchmarks should explicitly incorporate these or similar evidence-grounded metrics, especially for long-video question answering, dense video captioning, and generative reasoning tasks, so as to distinguish models that merely produce plausible language from those that generate temporally aligned and evidence-faithful responses.

5.6. Challenges and Future Directions

Several research directions follow from the above analysis. First, long-video understanding requires temporally efficient and evidence-sensitive neural computation. Instead of simply extending context length or uniformly increasing the number of sampled frames, future models should learn event-centric memory, adaptive frame or clip selection, and hierarchical temporal abstraction. A key bottleneck in current architectures is that global attention mechanisms tend to dilute fine-grained temporal boundaries when video length

increases. This is particularly important because sparse or uniform sampling may miss short actions, fast transitions, and causally important moments, thereby weakening temporal grounding even when the final language response appears plausible. We hypothesize that boundary-preserving temporal compression can improve long-range reasoning stability under fixed computational budgets.

Second, multimodal video understanding should give greater attention to audio-aware modeling. Many recent video-language systems still rely primarily on visual frames and text, while audio is often used only indirectly through ASR transcripts or omitted from the main modeling pipeline. A structural limitation is that most fusion architectures align modalities at a coarse temporal resolution, which makes it difficult to preserve fine-grained acoustic events such as prosody shifts or overlapping sound cues. However, speech prosody, speaker emotion, music, environmental sounds, and non-speech acoustic events can provide semantic and affective cues that are not always recoverable from visual frames or transcripts alone. Future models should therefore treat audio as a first-class modality and develop mechanisms for aligning acoustic events with visual actions, textual semantics, and temporal evidence. We hypothesize that explicit audio-visual event alignment improves robustness under partial modality missingness.

Third, generative video understanding needs stronger control over hallucination and evidence grounding. Large language decoders improve the fluency and flexibility of captions, summaries, and answers, but a key failure mode is the weak coupling between generated tokens and intermediate visual evidence representations. This risk is especially serious when relevant frames are not sampled, audio cues are ignored, or temporal evidence is ambiguous. Future systems should jointly localize visual, textual, and acoustic evidence, distinguish observed content from inferred or uncertain content, and provide calibrated confidence when evidence is incomplete or contradictory. A plausible direction is to explicitly constrain generation with retrieval or grounding modules that operate on structured temporal evidence.

Fourth, robustness under signal degradation should be more explicitly considered when comparing recent video understanding models. Real-world videos may contain visual noise, frame dropping, temporal truncation, ASR or OCR errors, subtitle mismatch, missing modalities, or audio–video desynchronization. A key bottleneck is that most current benchmarks evaluate models under clean and synchronized conditions, which hides sensitivity to modality-level corruption. These degradations may affect model families differently: video Transformers are vulnerable to corrupted or missing frames, video-language pretraining models may suffer from weakly aligned text-video pairs, Video-LLMs may hallucinate when fine-grained evidence is lost, and memory- or tool-augmented systems may propagate errors from retrieval, ASR, OCR, or external tools. Therefore, future benchmarks should report not only clean-set performance but also performance retention under degraded visual, temporal, textual, and cross-modal evidence, for example using relative robustness retention defined as $P_{\text{degraded}}/P_{\text{clean}}$.

Finally, evaluation should become capability-level and decomposable. A benchmark should enable asking whether an error arises from visual perception, audio understanding, temporal localization, cross-modal alignment, semantic generation, hallucination, or reasoning. A key limitation of current evaluation protocols is that they conflate these failure sources into a single score, making it difficult to diagnose model weaknesses. Such a decomposable evaluation framework would naturally align with the alignment-expression-reasoning structure proposed in this survey and would make reported progress across tasks more interpretable.

6. Conclusions

The development of multimodal video understanding remains fragmented across diverse task-specific directions and isolated benchmark settings. To overcome this limitation, this survey systematically analyzes state-of-the-art models and proposes a capability-based framework organized around alignment, expression, and reasoning. By organizing representative tasks, models, datasets, evaluation metrics, and benchmark results under this framework, the survey clarifies the internal relationship among cross-modal matching, semantic externalization, and evidence-based inference, thereby revealing the broader transition from task-specific architectures to large multimodal models. The comparative analysis indicates that recent progress is mainly driven by large-scale pretraining, fine-grained cross-modal interaction, temporal evidence modeling, and large-model-based generation, which jointly enhance the ability of models to represent complex video semantics and support flexible multimodal understanding. Future research should consider not only benchmark gains, but also scalable temporal modeling, robust multimodal fusion, evidence-grounded generation, and decomposable capability-level evaluation, thereby further improving the systematic development and practical evaluation of large multimodal video models.

Author Contributions: Conceptualization, R.Z.; methodology, D.P.; writing—original draft preparation, D.P.; writing—review and editing, R.Z., C.L. and Y.M.; validation, C.L., Z.Z., X.P. and Y.M.; supervision, R.Z.; project administration, R.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 72374154; the Key Project of the Ministry of Education of China, grant number 2024IT003; and the 20th Experimental Teaching Reform Special Foundation of Tongji University, grant number 0800104347.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The authors deeply appreciate the support associated with the funded projects listed above.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-Scale Video Classification with Convolutional Neural Networks. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2014; pp. 1725–1732. [CrossRef]
2. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 28th International Conference on Neural Information Processing Systems—Volume 1*; ACM: New York, NY, USA, 2014; pp. 568–576.
3. Xia, H.; Zhan, Y. A Survey on Temporal Action Localization. *IEEE Access* **2020**, *8*, 70477–70487. [CrossRef]
4. Abdar, M.; Kollati, M.; Kuraparthi, S.; Pourpanah, F.; McDuff, D.; Ghavamzadeh, M.; Yan, S.; Mohamed, A.; Khosravi, A.; Cambria, E.; et al. A Review of Deep Learning for Video Captioning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, 1–20. [CrossRef] [PubMed]
5. Jeshmol, P.J.; Kovoor, B.C. Video Question Answering: A survey of the state-of-the-art. *J. Vis. Commun. Image Represent.* **2024**, *105*, 104320. [CrossRef]
6. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2015.
7. Xu, P.; Zhu, X.; Clifton, D.A. Multimodal Learning with Transformers: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12113–12132. [CrossRef] [PubMed]

8. Nguyen, T.; Bin, Y.; Xiao, J.; Qu, L.; Li, Y.; Wu, J.Z.; Nguyen, C.D.; Ng, S.K.; Luu, A.T. Video-Language Understanding: A Survey from Model Architecture, Model Training, and Data Perspectives. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 3636–3657. [CrossRef]
9. Tang, Y.; Bi, J.; Xu, S.; Song, L.; Liang, S.; Wang, T.; Zhang, D.; An, J.; Lin, J.; Zhu, R.; et al. Video Understanding with Large Language Models: A Survey. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *36*, 1355–1376. [CrossRef]
10. Zhu, C.; Jia, Q.; Chen, W.; Guo, Y.; Liu, Y. Deep learning for video-text retrieval: A review. *Int. J. Multimed. Inf. Retr.* **2023**, *12*, 3. [CrossRef]
11. Zhang, H.; Sun, A.; Jing, W.; Zhou, J.T. Temporal Sentence Grounding in Videos: A Survey and Future Directions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10443–10465. [CrossRef] [PubMed]
12. Agrawal, M.; Jalal, A.S.; Sharma, H. A Review on VQA: Methods, Tools and Datasets. In *Proceedings of the 2023 International Conference on Computer Science and Emerging Technologies (CSET)*; IEEE: New York, NY, USA, 2023; pp. 1–6. [CrossRef]
13. Qasim, I.; Horsch, A.; Prasad, D. Dense Video Captioning: A Survey of Techniques, Datasets and Evaluation Protocols. *ACM Comput. Surv.* **2025**, *57*, 1–36. [CrossRef]
14. Laptev, I. On Space-Time Interest Points. *Int. J. Comput. Vis.* **2005**, *64*, 107–123. [CrossRef]
15. Dollar, P.; Rabaud, V.; Cottrell, G.; Belongie, S. Behavior recognition via sparse spatio-temporal features. In *Proceedings of the 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*; IEEE: New York, NY, USA, 2005; pp. 65–72. [CrossRef]
16. Wang, H.; Kläser, A.; Schmid, C.; Liu, C.L. Action recognition by dense trajectories. In *Proceedings of the CVPR 2011*; IEEE: New York, NY, USA, 2011; pp. 3169–3176. [CrossRef]
17. Brand, M.; Oliver, N.; Pentland, A. Coupled hidden Markov models for complex action recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 1997; pp. 994–999. [CrossRef]
18. Carreira, J.; Zisserman, A. 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)*; IEEE: New York, NY, USA, 2017; pp. 4724–4733. [CrossRef]
19. Bertasius, G.; Wang, H.; Torresani, L. Is Space-Time Attention All You Need for Video Understanding? In *Proceedings of the the 38th International Conference on Machine Learning*; PMLR: McKees Rocks, PA, USA, 2021; pp. 813–824.
20. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. ViViT: A Video Vision Transformer. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021; pp. 6816–6826. [CrossRef]
21. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video Swin Transformer. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 3192–3201. [CrossRef]
22. Tong, Z.; Song, Y.; Wang, J.; Wang, L. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. *arXiv* **2022**, arXiv:2203.12602. [CrossRef]
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2016; pp. 770–778. [CrossRef]
24. Hershey, S.; Chaudhuri, S.; Ellis, D.P.W.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B.; et al. CNN architectures for large-scale audio classification. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2017; pp. 131–135. [CrossRef]
25. Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; Plumbley, M.D. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2880–2894. [CrossRef]
26. Gong, Y.; Chung, Y.A.; Glass, J. AST: Audio Spectrogram Transformer. In *Proceedings of the Interspeech 2021*; ISCA: Red Hook, NY, USA, 2021; pp. 571–575. [CrossRef]
27. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Proceedings of the Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2020; pp. 12449–12460.
28. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186. [CrossRef]
29. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*; PMLR: McKees Rocks, PA, USA, 2021; pp. 8748–8763.
30. Baltrusaitis, T.; Ahuja, C.; Morency, L.P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [CrossRef] [PubMed]
31. Snoek, C.G.M.; Worring, M.; Smeulders, A.W.M. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*; ACM: New York, NY, USA, 2005; pp. 399–402. [CrossRef]

32. Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; Sun, C. Attention Bottlenecks for Multimodal Fusion. In *Proceedings of the Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2021; Volume 34, pp. 14200–14213.
33. Chen, W.; Wang, W.; Liu, L.; Lew, M.S. New Ideas and Trends in Deep Multimodal Content Understanding: A Review. *Neurocomputing* **2021**, *426*, 195–215. [CrossRef]
34. Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.H.; Chang, S.F.; Cui, Y.; Gong, B. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. In *Proceedings of the Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2021; pp. 24206–24221.
35. Xu, H.; Ghosh, G.; Huang, P.Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; Feichtenhofer, C. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 6787–6800. [CrossRef]
36. Bain, M.; Nagrani, A.; Varol, G.; Zisserman, A. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021; pp. 1708–1718. [CrossRef]
37. Wang, X.; Zhu, L.; Yang, Y. T2VLAD: Global-Local Sequence Alignment for Text-Video Retrieval. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021; pp. 5075–5084. [CrossRef]
38. Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; Wang, C. Learning the Best Pooling Strategy for Visual Semantic Embedding. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021; pp. 15784–15793. [CrossRef]
39. Li, X.; Zhou, F.; Xu, C.; Ji, J.; Yang, G. SEA: Sentence Encoder Assembly for Video Retrieval by Textual Queries. *IEEE Trans. Multimed.* **2021**, *23*, 4351–4362. [CrossRef]
40. Dzabraev, M.; Kalashnikov, M.; Komkov, S.; Petiushko, A. MDMMT: Multidomain Multimodal Transformer for Video Retrieval. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: New York, NY, USA, 2021; pp. 3349–3358. [CrossRef]
41. Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; Li, T. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing* **2022**, *508*, 293–304. [CrossRef]
42. Ma, Y.; Xu, G.; Sun, X.; Yan, M.; Zhang, J.; Ji, R. X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*; ACM: New York, NY, USA, 2022; pp. 638–647. [CrossRef]
43. Gorti, S.K.; Vouitsis, N.; Ma, J.; Golestan, K.; Volkovs, M.; Garg, A.; Yu, G. X-Pool: Cross-Modal Language-Video Attention for Text-Video Retrieval. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 4996–5005. [CrossRef]
44. Fu, T.J.; Li, L.; Gan, Z.; Lin, K.; Wang, W.Y.; Wang, L.; Liu, Z. VIOLET: End-to-End Video-Language Transformers with Masked Visual-token Modeling. *arXiv* **2022**, arXiv:2111.12681. [CrossRef]
45. Hu, P.; Huang, Z.; Peng, D.; Wang, X.; Peng, X. Cross-Modal Retrieval with Partially Mismatched Pairs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 9595–9610. [CrossRef] [PubMed]
46. Jin, L.; Li, Z.; Tang, J. Deep Semantic Multimodal Hashing Network for Scalable Image-Text and Video-Text Retrievals. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 1838–1851. [CrossRef] [PubMed]
47. Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. Intern VL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 24185–24198. [CrossRef]
48. Wu, Z.; Weng, Z.; Peng, W.; Yang, X.; Li, A.; Davis, L.S.; Jiang, Y.G. Building an Open-Vocabulary Video CLIP Model with Better Architectures, Optimization and Data. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *36*, 4747–4762. [CrossRef] [PubMed]
49. Wang, Z.; Gao, Z.; Yang, Y.; Wang, G.; Jiao, C.; Shen, H.T. Geometric Matching for Cross-Modal Retrieval. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, 5509–5521. [CrossRef] [PubMed]
50. Jiang, J.; Min, S.; Kong, W.; Wang, H.; Li, Z.; Liu, W. Tencent Text-Video Retrieval: Hierarchical Cross-Modal Interactions with Multi-Level Representations. *IEEE Access* **2025**, *13*, 122453–122463. [CrossRef]
51. Tu, R.C.; Ji, Y.; Jiang, J.; Kong, W.; Cai, C.; Zhao, W.; Wang, H.; Yang, Y.; Liu, W. Global and Local Semantic Completion Learning for Vision-Language Pre-Training. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 11065–11079. [CrossRef] [PubMed]
52. Wang, T.; Zhang, R.; Lu, Z.; Zheng, F.; Cheng, R.; Luo, P. End-to-End Dense Video Captioning with Parallel Decoding. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021; pp. 6827–6837. [CrossRef]
53. Lin, K.; Li, L.; Lin, C.C.; Ahmed, F.; Gan, Z.; Liu, Z.; Lu, Y.; Wang, L. SwinBERT: End-to-End Transformers with Sparse Attention for Video Captioning. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 17928–17937. [CrossRef]

54. Wang, T.; Zheng, H.; Yu, M.; Tian, Q.; Hu, H. Event-Centric Hierarchical Representation for Dense Video Captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 1890–1900. [CrossRef]
55. Zhao, B.; Gong, M.; Li, X. Hierarchical multimodal transformer to summarize videos. *Neurocomputing* **2022**, *468*, 360–369. [CrossRef]
56. Ye, H.; Li, G.; Qi, Y.; Wang, S.; Huang, Q.; Yang, M.H. Hierarchical Modular Network for Video Captioning. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 17918–17927. [CrossRef]
57. Li, L.; Gao, X.; Deng, J.; Tu, Y.; Zha, Z.J.; Huang, Q. Long Short-Term Relation Transformer with Global Gating for Video Captioning. *IEEE Trans. Image Process.* **2022**, *31*, 2726–2738. [CrossRef] [PubMed]
58. Ryu, H.; Kang, S.; Kang, H.; Yoo, C.D. Semantic Grouping Network for Video Captioning. *Proc. Aaai Conf. Artif. Intell.* **2021**, *35*, 2514–2522. [CrossRef]
59. Liu, F.; Ren, X.; Wu, X.; Yang, B.; Ge, S.; Sun, X. O2NA: An Object-Oriented Non-Autoregressive Approach for Controllable Video Captioning. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 281–292. [CrossRef]
60. Shen, Y.; Gu, X.; Xu, K.; Fan, H.; Wen, L.; Zhang, L. Accurate and Fast Compressed Video Captioning. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2023; pp. 15512–15521. [CrossRef]
61. Gu, X.; Chen, G.; Wang, Y.; Zhang, L.; Luo, T.; Wen, L. Text with Knowledge Graph Augmented Transformer for Video Captioning. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023; pp. 18941–18951. [CrossRef]
62. Yang, A.; Nagrani, A.; Seo, P.H.; Miech, A.; Pont-Tuset, J.; Laptev, I.; Sivic, J.; Schmid, C. Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023; pp. 10714–10726. [CrossRef]
63. He, B.; Li, H.; Jang, Y.K.; Jia, M.; Cao, X.; Shah, A.; Shrivastava, A.; Lim, S.N. MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 13504–13514. [CrossRef]
64. Zhou, X.; Arnab, A.; Buch, S.; Yan, S.; Myers, A.; Xiong, X.; Nagrani, A.; Schmid, C. Streaming Dense Video Captioning. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 18243–18252. [CrossRef]
65. Zhang, C.; Jian, Y.; Ouyang, Z.; Vosoughi, S. Pretrained Image-Text Models are Secretly Video Captioners. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 292–305. [CrossRef]
66. Wang, Y.; Liu, Y.; Zhou, S.; Huang, Y.; Tang, C.; Zhou, W.; Chen, Z. Emotion-Oriented Cross-Modal Prompting and Alignment for Human-Centric Emotional Video Captioning. *IEEE Trans. Multimed.* **2025**, *27*, 3766–3780. [CrossRef]
67. Song, P.; Guo, D.; Yang, X.; Tang, S.; Wang, M. Emotional Video Captioning with Vision-Based Emotion Interpretation Network. *IEEE Trans. Image Process.* **2024**, *33*, 1122–1135. [CrossRef] [PubMed]
68. Xi, Z.; Shi, G.; Sun, H.; Zhang, B.; Li, S.; Wu, L. EIKA: Explicit & Implicit Knowledge-Augmented Network for entity-aware sports video captioning. *Expert Syst. Appl.* **2025**, *274*, 126906. [CrossRef]
69. Xiong, Y.; He, C.; Chen, L.; Cai, T. Spatio-temporal graph interaction networks for teacher behavior description in classroom scene. *Eng. Appl. Artif. Intell.* **2025**, *159*, 111668. [CrossRef]
70. Jing, S.; Zhang, H.; Zeng, P.; Gao, L.; Song, J.; Shen, H.T. Memory-Based Augmentation Network for Video Captioning. *IEEE Trans. Multimed.* **2024**, *26*, 2367–2379. [CrossRef]
71. Liang, Y.; Zhu, L.; Wang, X.; Yang, Y. IcoCap: Improving Video Captioning by Compounding Images. *IEEE Trans. Multimed.* **2024**, *26*, 4389–4400. [CrossRef]
72. Park, J.; Lee, J.; Sohn, K. Bridge to Answer: Structure-aware Graph Interaction Network for Video Question Answering. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021; pp. 15521–15530. [CrossRef]
73. Xiao, J.; Zhou, P.; Yao, A.; Li, Y.; Hong, R.; Yan, S.; Chua, T.S. Contrastive Video Question Answering via Video Graph Transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13265–13280. [CrossRef] [PubMed]
74. Wang, T.; Hou, B.; Li, J.; Shi, P.; Zhang, B.; Snoussi, H. TASTA: Text-Assisted Spatial and Temporal Attention Network for Video Question Answering. *Adv. Intell. Syst.* **2023**, 2200131. [CrossRef]
75. Liu, J.; Wang, G.; Xie, J.; Zhou, F.; Xu, H. Video Question Answering with Semantic Disentanglement and Reasoning. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 3663–3673. [CrossRef]
76. Gao, D.; Zhou, L.; Ji, L.; Zhu, L.; Yang, Y.; Shou, M.Z. MIST: Multi-modal Iterative Spatial-Temporal Transformer for Long-form Video Question Answering. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023; pp. 14773–14783. [CrossRef]

77. Xiao, J.; Shang, X.; Yao, A.; Chua, T.S. NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021; pp. 9772–9781. [CrossRef]
78. Li, Y.; Wang, X.; Xiao, J.; Ji, W.; Chua, T.S. Invariant Grounding for Video Question Answering. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 2918–2927. [CrossRef]
79. Zang, C.; Wang, H.; Pei, M.; Liang, W. Discovering the Real Association: Multimodal Causal Reasoning in Video Question Answering. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023; pp. 19027–19036. [CrossRef]
80. Yang, X.; Zeng, J.; Guo, D.; Wang, S.; Dong, J.; Wang, M. Robust video question answering via contrastive cross-modality representation learning. *Sci. China Inf. Sci.* **2024**, *67*, 202104. [CrossRef]
81. Zhang, T.; Zhu, Y.; Wu, B.; Zheng, C.; Tan, J.; Xiong, Z. A general debiasing framework with counterfactual reasoning for multimodal public speaking anxiety detection. *Neural Netw.* **2025**, *187*, 107314. [CrossRef] [PubMed]
82. Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T.L.; Bansal, M.; Liu, J. Less is More: CLIPBERT for Video-and-Language Learning via Sparse Sampling. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021; pp. 7327–7337. [CrossRef]
83. Zellers, R.; Lu, X.; Hessel, J.; Yu, Y.; Park, J.S.; Cao, J.; Farhadi, A.; Choi, Y. MERLOT: Multimodal Neural Script Knowledge Models. In *Proceedings of the Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2021; pp. 23634–23651.
84. Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; Schmid, C. Zero-shot video question answering via frozen bidirectional language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2022; pp. 124–141.
85. Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; Schmid, C. Learning to Answer Visual Questions From Web Videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 3202–3218. [CrossRef] [PubMed]
86. Weng, Y.; Han, M.; He, H.; Chang, X.; Zhuang, B. LongVLM: Efficient Long Video Understanding via Large Language Models. In *Proceedings of the Computer Vision—ECCV 2024*; Springer: Berlin/Heidelberg, Germany, 2025; pp. 453–470. [CrossRef]
87. Min, J.; Buch, S.; Nagrani, A.; Cho, M.; Schmid, C. MoReVQA: Exploring Modular Reasoning Models for Video Question Answering. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 13235–13245. [CrossRef]
88. Yu, T.; Fu, K.; Wang, S.; Huang, Q.; Yu, J. Prompting Video-Language Foundation Models with Domain-specific Fine-grained Heuristics for Video Question Answering. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 1615–1630. [CrossRef]
89. Xiao, J.; Huang, N.; Qin, H.; Li, D.; Li, Y.; Zhu, F.; Tao, Z.; Yu, J.; Lin, L.; Chua, T.S.; et al. VideoQA in the Era of LLMs: An Empirical Study. *Int. J. Comput. Vis.* **2025**, *133*, 3970–3993. [CrossRef]
90. Xu, J.; Mei, T.; Yao, T.; Rui, Y. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2016; pp. 5288–5296. [CrossRef]
91. Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.F.; Wang, W.Y. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2019; pp. 4580–4590. [CrossRef]
92. Miech, A.; Zhukov, D.; Alayrac, J.B.; Tapaswi, M.; Laptev, I.; Sivic, J. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2019; pp. 2630–2640. [CrossRef]
93. Yao, L.; Torabi, A.; Cho, K.; Ballas, N.; Pal, C.; Larochelle, H.; Courville, A. Describing Videos by Exploiting Temporal Structure. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2015; pp. 4507–4515. [CrossRef]
94. Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; Niebles, J.C. Dense-Captioning Events in Videos. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2017; pp. 706–715. [CrossRef]
95. Bagher Zadeh, A.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 2236–2246. [CrossRef]
96. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 527–536. [CrossRef]

97. Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; Kim, G. TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2017; pp. 1359–1367. [CrossRef]
98. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; pp. 311–318. [CrossRef]
99. Banerjee, S.; Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2005; pp. 65–72.
100. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Text Summarization Branches out*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2004; pp. 74–81.
101. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based image description evaluation. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2015; pp. 4566–4575. [CrossRef]
102. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the International Conference on Learning Representations*; Curran Associates, Inc.: Red Hook, NY, USA, 2020.
103. Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; Choi, Y. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*; Moens, M.F., Huang, X., Specia, L., Yih, S.W.t., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 7514–7528. [CrossRef]
104. Zhao, H.; Chen, Z.; Yang, Y. Multi-scale features with temporal information guidance for video captioning. *Eng. Appl. Artif. Intell.* **2024**, *137*, 109102. [CrossRef]
105. Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; Schmid, C. Just Ask: Learning to Answer Questions from Millions of Narrated Videos. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021; pp. 1666–1677. [CrossRef]
106. Yuan, Y.; Lan, X.; Wang, X.; Chen, L.; Wang, Z.; Zhu, W. A Closer Look at Temporal Sentence Grounding in Videos: Dataset and Metric. In *Proceedings of the 2nd International Workshop on Human-Centric Multimedia Analysis*; ACM: New York, NY, USA, 2021; HUMA'21; pp. 13–21. [CrossRef]
107. Li, C.; Im, E.W.; Fazli, P. VidHalluc: Evaluating Temporal Hallucinations in Multimodal Large Language Models for Video Understanding. In *Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2025; pp. 13723–13733. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.