A Survey on Hallucination in Video Understanding: Taxonomy, Causes, and Mitigation Techniques

Anonymous authors
Paper under double-blind review

Abstract

Video Large Language Models (Vid-LLMs) have recently achieved strong performance across a wide range of video understanding tasks, including question answering, captioning, and multimodal reasoning. However, these models frequently produce outputs that are not faithfully grounded in the underlying video content, a phenomenon commonly referred to as hallucination. Compared with hallucination in text-only or image-based models, hallucination in video understanding is further complicated by temporal dynamics, motion interpretation, long-context dependencies, and event-level reasoning. In this survey, we present a comprehensive review of hallucination in Vid-LLMs. We begin with a unified taxonomy, which categorizes different hallucination phenomena, and then organize existing mitigation strategies according to the failure mechanisms they address. We also discuss key open challenges and outline some promising research directions.

1 Introduction

AI models are increasingly deployed in real-world applications requiring perception, reasoning, and decision-making at scale. Video understanding has emerged as a core task for modern multimodal foundation models. These systems take dynamic visual streams as input, recognize objects and actions, track interactions, and reason about events over time. Video understanding systems are now widely adopted in high-stakes settings, including but not limited to: surveillance and detection, autonomous driving, human-robot interactions, and large-scale video content moderation (Chen et al., 2024b; Eze & Crick, 2025; Wang et al., 2025q).

Incorrect or unsupported interpretations of video content, which are commonly referred to as **hallucinations**, can result in degraded system performance and serious safety, legal, or ethical consequences (Wang et al., 2025m). Nonexistent actions, misidentified objects, or fabricated event explanations may propagate to downstream decision-makers, amplifying their impact beyond isolated errors (Wang et al., 2025o). Rather than viewing hallucination in video understanding merely as a modeling imperfection, we should recognize it as a system-level reliability issue that directly affects trustworthiness in deployment. This motivates the need for principled understanding and mitigation of video hallucination as a prerequisite for building dependable video understanding systems.

Prior surveys have established hallucination as a well-defined research problem for large language models (LLMs) in the text domain. Works such as (Huang et al., 2025b; Tonmoy et al., 2024) have provided principled taxonomies, analyzed failure mechanisms, and summarized mitigation strategies. There are also existing surveys on hallucination in large vision-language models (VLMs) for the image domain, including categorization of hallucination behaviors and representative mitigation methods (Liu et al., 2024a; Bai et al., 2024). However, extending these established perspectives to video understanding is non-trivial, as video introduces temporal and event-level challenges that lead to unique hallucination behaviors.

Video understanding goes beyond image-based perception: it requires temporal modeling, event-level reasoning, and long-range information integration across largely many frames. While video large language models (Vid-LLMs) certainly inherit certain perception-level hallucinations from image-based VLMs (e.g., object existence, attributes, and spatial relations), many failures are inherently video-specific. For example, models may mishandle temporal understanding, misread motion and action dynamics, or lose information under

long-context memory and retrieval even when single-frame perception is correct (Chang et al., 2025b; Tang et al., 2025f). Video understanding frequently involves higher-level reasoning over events, including causality, intent, and multi-hop logic, where models may produce confident yet unsupported explanations (Jiang et al., 2025b).

Despite substantial recent efforts devoted to advancing Vid-LLMs (Tang et al., 2025c; Nguyen et al., 2024), their hallucinations have not yet been systematically reviewed. Existing surveys on video understanding largely organize the literature around model architectures, training paradigms, datasets, and benchmark performance (Nguyen et al., 2024; Tang et al., 2025c; Zhong et al., 2022; Zhou et al., 2024; Madan et al., 2024; Zou et al., 2024), so they write about what video models can do and how to improve task-level accuracy. In those works, failure cases are typically discussed as general challenges rather than being formalized with distinct types, causes, and targeted remedies. A few cross-modal hallucination surveys have covered systems with video processing, but they distribute attention across multiple modalities and thus provide only coarse analyses (Sahoo et al., 2024; Xia et al., 2025). The literature lacks a unified vocabulary and structure for describing video hallucination and for connecting specific hallucination types to appropriate mitigation strategies. This survey fills this gap.

In this survey, we provide a unified, video-centric taxonomy of hallucination, connect each hallucination category to corresponding mitigation techniques, and identify key open challenges and future research directions. By structuring prior work in this way, the survey enables both researchers and practitioners to reason about video understanding failures in a principled manner and to develop mitigation approaches that address specific hallucination mechanisms and real-world deployment scenarios.

The contributions of this survey include:

- **Comprehensive taxonomy of video understanding hallucination.** We provide six major categories: (1) temporal understanding hallucination, (2) motion and action dynamics hallucination, (3) long-context video understanding hallucination, (4) object hallucination, (5) causal and reasoning hallucination, and (6) cross-modal hallucination.
- **Systematic categorization of mitigation techniques.** We systematically review and organize recent mitigation methods for video understanding hallucination, aligning mitigation strategies with specific hallucination categories.
- **Forward-looking research directions toward trustworthy Vid-LLMs.**

  Beyond summarizing existing methods, we outline some promising directions for improving Vid-LLM robustness. We highlight the need for hallucination-specific benchmarks that measure grounding fidelity rather than relying solely on task accuracy, discuss out-of-schema strategies that enable models to handle novel entities and event compositions, and explore world-modeling approaches for temporal causality and action dynamics.

2 Hallucination in the Era of Video Understanding

2.1 Video Large Language Models

Vid-LLMs are multimodal foundation models that integrate video perception with LLMs to perform video understanding and reasoning tasks under natural language instructions. These tasks include video question answering, temporal grounding, video captioning, and causal inference (Tang et al., 2025c).

Vid-LLMs extend VLMs from static images to dynamic video inputs. While VLMs encode a single or a few images into visual tokens aligned with a language model, Vid-LLMs must process frame sequences or clips and model their temporal dependencies (Liu et al., 2024a).

Most Vid-LLMs consist of three components: (i) *a video encoder* that extracts spatio-temporal representations from video frames, (ii) *a temporal aggregation or memory mechanism* that compresses and organizes long video streams, and (iii) *a large language model* that performs reasoning and generation conditioned on the encoded video content (Nguyen et al., 2024).

Vid-LLMs face substantially higher complexity than image-based VLMs. Videos are orders of magnitude longer than images in both input size and semantic scope, often containing multiple events, evolving states, and long-range dependencies. Vid-LLMs must not only recognize objects and attributes, but also understand temporal order, motion patterns, event boundaries, and cross-event relations (Sahoo et al., 2024). These requirements fundamentally change both the failure modes and reliability properties of the model.

2.2 Hallucination in Vid-LLMs

In Vid-LLMs, hallucination refers to generated content that lacks faithful grounding in the actual video input. A hallucination can occur when a Vid-LLM produces descriptions, answers, or explanations that contradict or are unsupported by the visual-temporal evidence present in the video.

This definition parallels hallucination in image-based large VLMs, where hallucination is defined as misalignment between image input and textual output (Liu et al., 2024a). Clearly, hallucination in Vid-LLMs can exhibit fundamentally different characteristics due to the temporal and sequential nature of video data. This complexity arises from the need for Vid-LLMs to reason over multiple frames and, in many cases, across extended clips (Tang et al., 2025c).

In particular, a Vid-LLM may correctly recognize objects in individual frames yet hallucinate how these objects move, interact, or change over time. The model may invent actions that never occur, misorder events, attribute wrong causality, or confidently summarize events that are absent from the video. Also, in long videos with minutes or even hours of visual content, models may forget earlier events, merge attributes from different temporal segments, or retrieve incorrect or incomplete memories when answering downstream queries.

These distinctions indicate that hallucination in video understanding requires a dedicated formulation that explicitly accounts for temporal dynamics, motion, and long-context dependencies.

2.3 Unique Challenges regarding Hallucination in Vid-LLMs

Hallucination in Vid-LLMs presents challenges far more complex than those encountered in static image understanding.

Temporal complexity and ambiguity. Videos encode information across time, where meaning depends on event order, duration, and temporal context. When models develop ambiguous temporal understanding, they infer incorrect actions or event sequences. Small temporal misalignments can propagate into large semantic hallucinations at the event level.

Long-context compression and memory degradation. Vid-LLMs may not directly process all frames in a video sequence. Instead, they rely on sampling and compression to fit input into limited tokens or memory slots to enable long-context modeling in practice. But this inevitably discards fine-grained video details. As video length increases, models forget early events, conflate attributes across distant segments, or hallucinate on details.

Entanglement of different errors. In Vid-LLMs, hallucinations may stem from incorrect or incomplete object recognition, which then propagates through downstream temporal and linguistic reasoning modules. But, even when frame-level object recognition is accurate, errors in temporal aggregation, cross-clip reasoning, or causal inference can lead to wrong conclusions, too. Therefore, perceptual, temporal, and reasoning errors are possibly entangled, which makes it challenging to determine where a hallucination originates.

3 Hallucination Taxonomy

To systematically analyze Vid-LLMs' hallucination, we organize existing failure modes into a unified taxonomy. Figure 3 groups hallucination phenomena into six major categories. Each category captures distinct reliability failures that can arise while other aspects of video perception remain accurate.

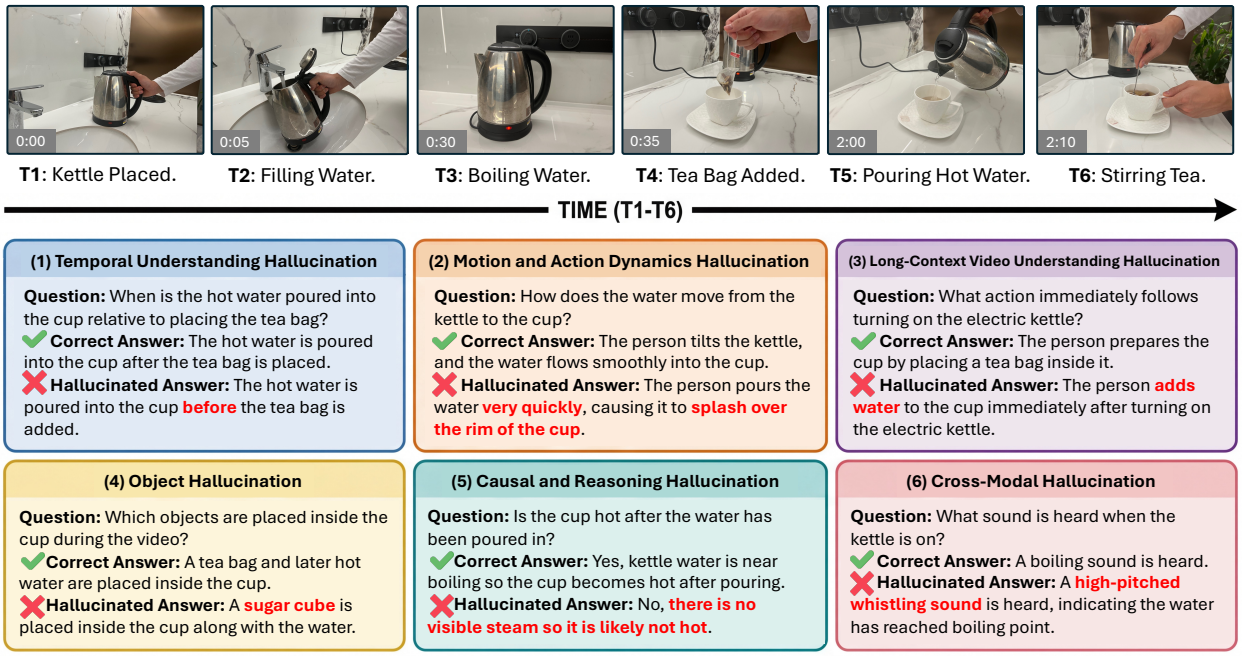

Figure 1: **Simple illustration of major video-understanding hallucination types.** Using a tea-making example with a timeline of key moments (T1–T6), the figure contrasts correct and hallucinated answers across six categories: (1) temporal understanding hallucination, (2) motion and action dynamics hallucination, (3) long-context video understanding hallucination, (4) object hallucination, (5) causal and reasoning hallucination, and (6) cross-modal hallucination.

Figures 1 and 2 demonstrate different hallucination categories through concrete examples.

3.1 Temporal Understanding Hallucination

Temporal understanding hallucination is the failure where a Vid-LLM produces incorrect temporal ordering or event timing. The model may correctly recognize which events are present but misrepresent their chronological order, relative timing, or durations. Prior work on temporal reasoning and moment-level understanding has shown that such temporal errors remain prevalent (Ahn et al., 2025; Huang et al., 2024a). We identify two major manifestations of this category: *sequencing errors* and *temporal grounding errors*.

Sequencing Errors. Sequencing errors arise when a model messes up the chronological order of events while correctly identifying the individual events themselves (Wu et al., 2025a; Plizzari et al., 2025; Li et al., 2025a). A common example is "the person sits down and then stands up" when the video actually shows standing and then sitting. Another case is "the player celebrates and then scores" when the score clearly precedes the celebration.

Empirical analyses of temporal ordering have shown that models frequently struggle with maintaining consistent event order, leading to swapped or reversed narratives in generated descriptions (Ahn et al., 2025; Xue et al., 2025b).

Temporal Grounding Errors. Temporal grounding errors arise *when* an event is claimed to happen along the timeline but assigned incorrect timestamps or boundaries. For instance, stating that "a person starts running" at 0:10 when it actually starts at another time, or describing that "the accident happens in the last few seconds" when it occurs near the middle of the video (Guo et al., 2025d; Chen et al., 2024e). Another common symptom is over-extending event duration, where a brief hand gesture is described as spanning an entire segment. These errors have been widely documented in fine-grained grounding and

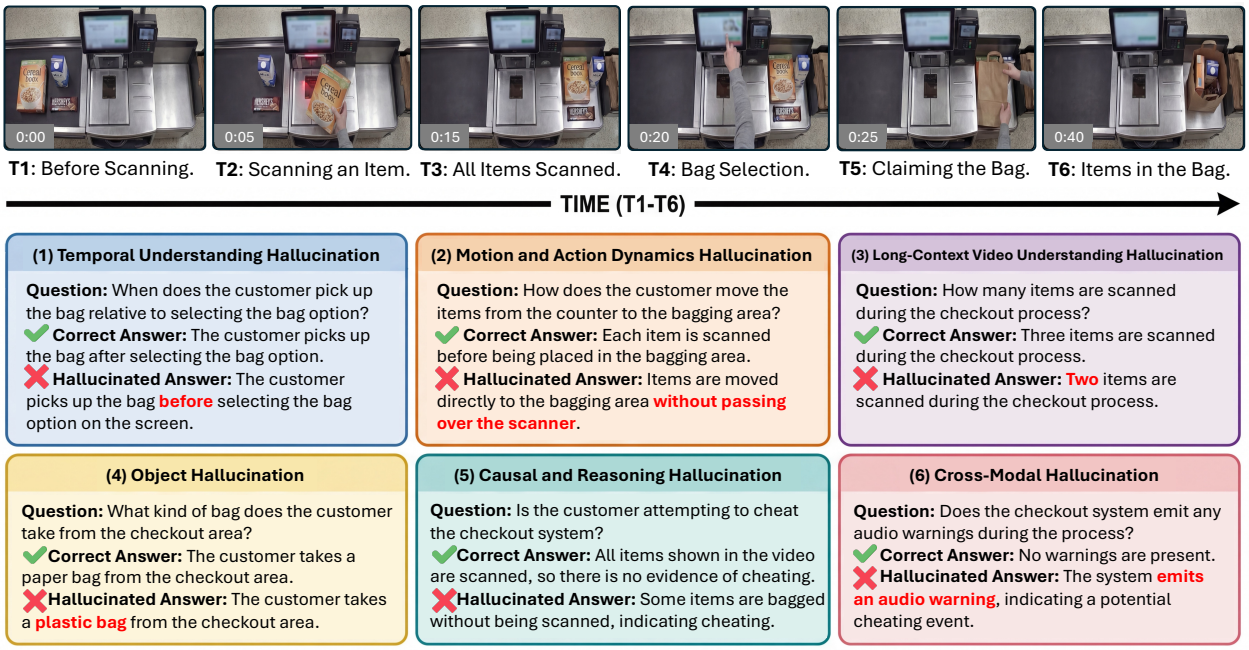

Figure 2: **Real-world example of major video-understanding hallucination types.** Using a grocery store self-checking example with a timeline of key moments (T1–T6), the figure contrasts correct and hallucinated answers across six categories: (1) temporal understanding hallucination, (2) motion and action dynamics hallucination, (3) long-context video understanding hallucination, (4) object hallucination, (5) causal and reasoning hallucination, and (6) cross-modal hallucination.

moment localization tasks (Huang et al., 2024a; Wang et al., 2024b; Wu et al., 2025e; Pramanick et al., 2025).

3.2 Motion and Action Dynamics Hallucination

Motion and action dynamics hallucination refers to failures in capturing how movements unfold over time, which is important when to distinguishing between actions relies on subtle temporal cues or when correct understanding requires consistent modeling of physical trajectories. Such hallucinations arise when models fail to encode fine-grained inter-frame changes, causing them to hallucinate motion that never occurred or misinterpret visually similar actions. We highlight two recurring error patterns: *fine-grained action errors* and *physics attribute errors*.

Fine-Grained Action Errors. Even when the model understands the scene, it can fail to distinguish subtle motions (Du et al., 2025b; Hong et al., 2025b). For example, describing a person as running when they are actually walking quickly, or claiming that a person is rubbing their eyes when they are merely adjusting their glasses. The literature attributes it to insufficient sensitivity to small-amplitude temporal differences across adjacent frames (Shao et al., 2020; Liu et al., 2022b).

Physics Attribute Errors. Physics attribute errors are failures in tracking the kinematic evolution of objects (Wu et al., 2025b). Typical mistakes include stating that a ball rolls to the left when it does to the right, or stating the speed wrong. Models may also hallucinate physically implausible transitions, such as sudden direction changes without a transition. Empirical studies such as PhyVLLM (Zhan et al., 2025b) and PiTe (Liu et al., 2024e) claim that incorrect reasoning about motion direction, trajectory, and velocity remains common in Vid-LLMs.

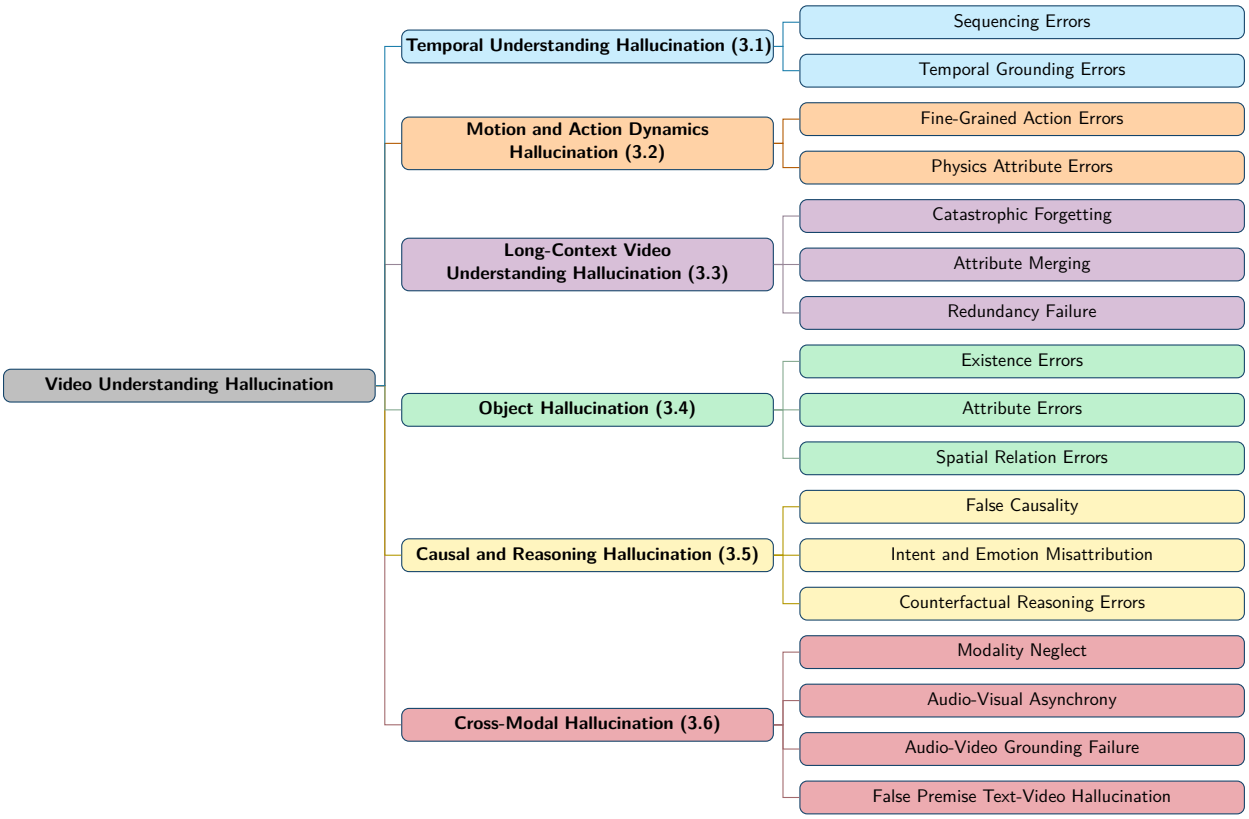

Figure 3: Taxonomy of video understanding hallucination, organized by category and error type.

3.3 Long-Context Video Understanding Hallucination

Long-context video understanding hallucination arises as its name suggests. When models integrate evidence across thousands to millions of frames, constraints from limited context windows, compressed representations, sliding windows, or limited external memory can distort what the model recalls or attends to over time. We identify three characteristic error patterns: *catastrophic forgetting*, *attribute merging*, and *redundancy failure*.

Catastrophic Forgetting. Earlier evidence tends to gradually disappear as the model processes later frames (Xiong et al., 2025; Wang et al., 2024p). Catastrophic forgetting occurs when a model correctly had encoded an early event or fact but later fails to recall it or produces contradictory statements (Jin et al., 2025d; Yang et al., 2025d; Li et al., 2024d; Yuan et al., 2025b). Consider an hour-long lecture video where a model correctly recognizes that the instructor taught something in the first few minutes, yet later claims it didn't happen when queried near the end.

Attribute Merging. Attribute merging refers to information *conflation*, i.e., the model incorrectly blends attributes of objects, people, or scenes that appear at different times (Lu et al., 2025). For example, after observing a red car early and a blue car later, the model incorrectly answers with a red car in the later scene.

Redundancy Failure. Long videos often contain large volumes of repetitive or low-information frames, which can suppress attention to sparse but critical evidence (Yunzhuzhang et al., 2025; Hu et al., 2025a; Yu et al., 2024a). Redundancy failure arises when models become overwhelmed by static scenes or slow-changing environments, causing salient events to get diluted, thus triggering hallucinated details (Zhang et al., 2025f;b; Tan et al., 2026). In an hour-long security camera feed with mostly event-free frames, a person may briefly walk through the hallway mid-video, yet the model fails to identify or report this event

due to the dominance of redundant low-activity content. The signal-to-noise ratio becomes particularly problematic in such scenarios.

3.4 Object Hallucination

Object hallucination describes perception-level failures where a model cannot reliably extract object-centric evidence and consequently generates unsupported or inconsistent claims. Compared to static images, object understanding in videos presents greater challenges, as objects may undergo motion, occlusion, viewpoint changes, and appearance variation. Hallucination may arise in identifying *what* objects exist, *which* attributes they exhibit, or *how* they are spatially arranged in a dynamic video setting. We summarize three representative forms of hallucination: *existence errors*, *attribute errors*, and *spatial relation errors*.

Existence Errors. Existence errors are hallucinated objects that do not appear in the video. Temporal sparsity and occlusion often exacerbate such failures, where objects may be visible only briefly or partially, making consistent presence tracking difficult (Lusha et al., 2025; Rawal et al., 2025; Wang et al., 2024n). Consider a cooking video that shows only a person stirring a pot yet a model may hallucinate the presence of a knife even though no knife ever appears. This type of unsupported entity generation has been reported in (Chang et al., 2025b; Wang et al., 2025b).

Attribute Errors. Attribute errors refer to misidentified visual properties, such as color, shape, size, or textual content, often due to lighting variation, motion blur, and viewpoint shifts (Rawal et al., 2025; Patraucean et al., 2023; Yang et al., 2025a). The model may correctly recognize a car but incorrectly claim it is red when it is actually blue. Such errors frequently occur in long-video understanding, where incomplete or compressed representations amplify the model's confusion (Wang et al., 2024i; Zou et al., 2024).

Spatial Relation Errors. Spatial relation errors are incorrectly described configurations or relative positions of objects. For example, it may state "the cup is on the table" when the video shows the cup being held in a hand. These errors may be attributed to the changing spatial groundings like object motion and camera movement (Wang et al., 2024n). Imperfect spatial grounding may cause models to generate plausible but unsupported relations (Wang et al., 2025d; Ouyang et al., 2025a; Tang et al., 2025c).

3.5 Causal and Reasoning Hallucination

Causal and reasoning hallucination arises when a model produces logically unsupported conclusions on top of otherwise correct perception. Here, models hallucinate unobservable concepts such as motivations, emotions, intentions, or causal relations. Unlike object hallucination (e.g., hallucinating an object), this category reflects breakdowns in higher-level inference over events and interactions (Jiang et al., 2025b). Such failures are particularly salient in video understanding because correct reasoning often requires integrating temporal cues, social context; and multi-step event structures. We identify three representative forms: *false causality*, *intent and emotion misattribution*, and *counterfactual reasoning errors*.

False Causality. False causality is a spurious cause-to-effect link: the model correctly recognizes events but constructs an unsupported causal explanation often by relying on stereotypical correlations instead of video evidence (Guda et al., 2024; Li et al., 2022b; Chen et al., 2024f). For example, seeing a video where a man suddenly starts running, the model may claim he is running because someone is chasing him. Models often conflate correlation with causation and over-rely on commonsense priors (Liu et al., 2023c; Wei et al., 2023a).

Intent and Emotion Misattribution. Intent and emotion misattribution is the *over-interpretation of social evidence*: the model maps observable behavior into mental-state labels without sufficient visual support (Li et al., 2023a; Ye et al., 2024; Peng et al., 2025; Chen et al., 2025k; Kong et al., 2025). For instance, a model may claim that "two people are arguing" based on their gestures, when the video actually shows them laughing with gestures.

Counterfactual Reasoning Errors. Counterfactual reasoning errors emerge under *hypothetical intervention* queries, where correctness requires reasoning about alternative outcomes consistent with rules and causal structure (Chen et al., 2025h; Huang et al., 2025d; Poppi et al., 2026) indicated by the video. For example, when asked "What would happen if the player missed the shot?", the model may hallucinate an outcome that contradicts the rules of the game played in the video.

3.6 Cross-Modal Hallucination

Cross-modal hallucination arises when models incorrectly integrate multimodal evidence (e.g., vision, audio, and text), producing outputs that violate cross-modal consistency (Guo et al., 2025f; Shen et al., 2023). We categorize it into four recurring patterns: *modality neglect*, *audio-visual asynchrony*, *audio-video grounding failure*; and *false premise text–video hallucination*.

Modality Neglect. Modality neglect occurs when one modality (typically vision) dominates, causing the model to ignore another (typically audio), even when it is salient (Guo et al., 2025f; Shen et al., 2023; Li et al., 2025k). For example, a video shows a person smiling and nodding while the audio contains another person saying "I feel terrible and I need help", but a model that discards the audio clues may output "a person is listening and agreeing" (Hanna-Asaad et al., 2024; Cheng et al., 2024b).

Audio-Visual Asynchrony. Audio-visual asynchrony is a temporal alignment failure: the model uses both modalities but binds auditory evidence to the wrong visual moment (Goel et al., 2024b; Tang et al., 2025d). For instance, it may describe a door slamming sound as occurring when a person merely approaches the door, rather than when the door actually closes.

Audio-Video Grounding Failure. Audio-video grounding failure concerns *source attribution*: the model assigns an audio signal to an incorrect visual entity or invents a source not responsible for the sound (Li et al., 2022a; Jiang & Yin, 2023). For example, in a scene with multiple people on screen and background television audio, the model may hallucinate that one of the visible individuals is speaking, rather than the off-screen TV (Li et al., 2024e).

False Premise Text-Video Hallucination. False premise text-video hallucination arises when the textual input embeds an assumption that conflicts with the video, and the model prioritizes linguistic grounding over visual grounding (Gao et al., 2025b). For example, when asked about the color of a cat in a video that contains no cat, the model may answer "the cat is white," hallucinating to satisfy the textual premise (Seth et al., 2025).

4 Foundations of Hallucination Mitigation

Before summarising specific mitigation techniques in Section 5 below, we review recurring methods underlying recent hallucination mitigation research. As hallucination arises from failures at different stages of the Vid-LLM pipeline, existing mitigation approaches can be understood as interventions at the corresponding stages. The following subsections review these methods, including reinforcement learning (RL) and preference optimization, supervision augmentation with synthetic data, structured representations, long-context video evidence management, and inference-time interventions, organized by where they intervene (see Figure 4).

4.1 Reinforcement Learning and Preference Optimization

RL and preference optimization align LLMs with desired behaviors by introducing training objectives beyond standard likelihood-based learning (Zhang et al., 2025e; Rafailov et al., 2023). The language model is treated as a policy, and an external training signal evaluates generated outputs and updates the model parameters so that preferred responses become more likely (Liu et al., 2025c).

In RL-based methods, the model generates candidate outputs and receives reward signals measuring factual correctness, reasoning quality, or human preference alignment. Policy optimization then updates the model

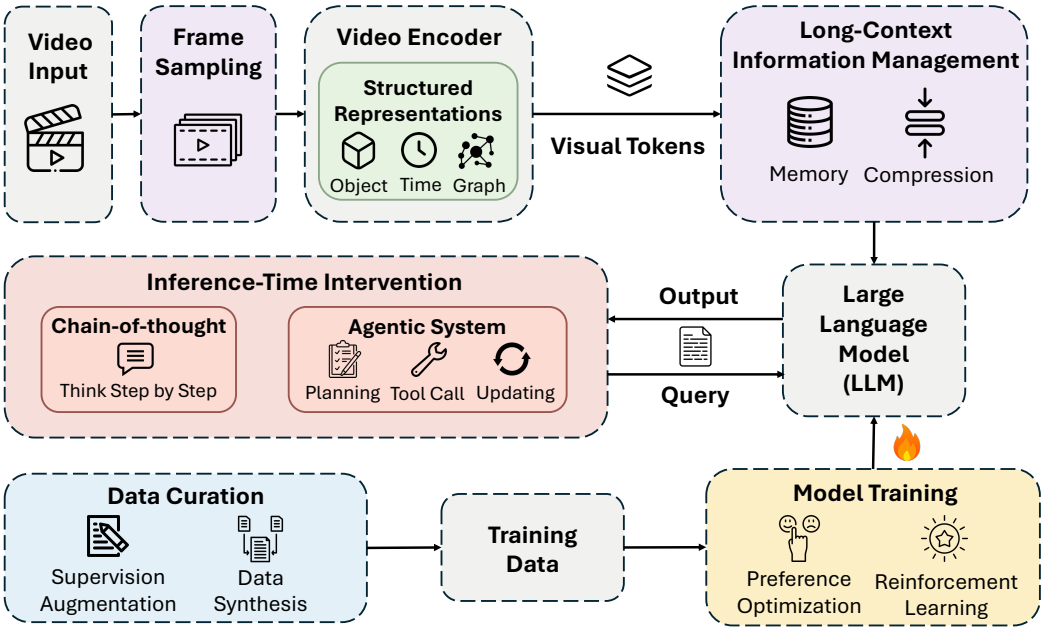

Figure 4: Illustration of representative hallucination mitigation strategies in Vid-LLMs organized by the stage of the pipeline where they intervene. Colored components highlight different methodological directions discussed in Section 4, with each color corresponding to one direction, while the grey blocks depict the standard Vid-LLMs pipeline.

to reinforce higher-reward responses and suppress lower-reward ones. For Vid-LLMs, effective reward design relies on verifiable supervision tied to video-specific structure, such as timestamps, segment boundaries, object trajectories, and event sequences, rather than generic answer correctness alone (Feng et al., 2025; Chen, 2025).

Preference optimization instead trains on pairs of candidate responses where one is preferred over the other (Huang et al., 2025a). In video-language settings, many hallucinations arise from subtle temporal or spatial misalignments rather than completely wrong answers, so preference construction benefits from fine-grained temporal or motion-aware supervision rather than relying on coarse human judgments (Li et al., 2025f; Kulkarni & Fazli, 2025).

4.2 Supervision Augmentation and Synthetic Data

Standard video datasets emphasize coarse descriptions and contain little supervision for subtle temporal relations, object interactions, or causal reasoning. Models may therefore never encounter training examples that explicitly penalize hallucinations of these kinds.

Synthetic supervision addresses this by constructing additional training examples that highlight hallucination-prone scenarios. Synthetic pipelines perturb or recompose video-text pairs to train the model on correct evidence grounding, for instance, by shuffling or reversing event sequences (Chen et al., 2023b; Vani et al., 2025), and concatenating shorter clips to simulate long videos (Ren et al., 2024b).

However, synthetic supervision introduces a trade-off between controllability and realism: artificially constructed examples may not reflect natural video distributions. The additional fine-grained temporal signals also increase memory and token requirements, so the effectiveness of synthetic data depends on both the quality of generated supervision and whether the model architecture can preserve these signals during long-context processing.

4.3 Structured Representations

Structured representations modify models to keep entities, relations, and interactions explicit during reasoning, rather than mapping visual inputs directly into dense token embeddings. Such models may benefit from intermediate, interpretable structures like relational graphs, modular feature groups, and entity-based representations (Mao et al., 2022).

In video understanding, hallucinations can be attributed to planning errors, not only perception errors, for example, answering before it has localized the relevant moments or decomposed the reasoning path. Structured representations help because they force intermediate grounding decisions before the final answer. An effective design to address the issue links observations across multiple frames into temporally coherent units through temporal linking, trajectory aggregation, or timestamp-aware embeddings; they encode both appearance features and temporal indices (Guo et al., 2025c; Wang et al., 2024a).

Another approach organizes video information into scene graphs or dynamic interaction graphs, representing entities as nodes and their spatial or temporal interactions as edges (Du & Wang, 2025). Reasoning over explicit relationships helps prevent the model from introducing unsupported objects, relations, or interactions (Huang et al., 2025c; Chu et al., 2025).

In practice, structured representations alone are insufficient. Their effectiveness depends on how the model uses them, which require trainings or agent pipelines that focus attention on query-relevant relations and evidence (Malik et al., 2025).

4.4 Long-Context Information Management

Long videos introduce a fundamental challenge: a single video may contain thousands of frames, while the model can process only a limited number of visual tokens (Xu et al., 2025c). Models must therefore selectively retain, compress, and retrieve information from earlier parts of the video.

Memory-based approaches augment the model with external or persistent memory that stores compact representations of previously processed segments, such as learned embeddings, key-value representations, and segment-level summaries. At query time, relevant memory entries are retrieved through similarity matching (Song et al., 2023; Jin et al., 2025c).

Token compression methods reduce visual tokens by merging redundant tokens using clustering, similarity-based merging, or learned compressors, fitting longer videos within fixed context budgets (Weng et al., 2024; Guo et al., 2025b). Empirically, effective compression preserves the spatial and temporal structure of events rather than minimizing token counts uniformly, because collapsing distinct moments into indistinguishable tokens can lead to hallucinated actions or causal relations (Sun et al., 2025a).

Hallucination in long videos can also be a retrieval failure disguised as a reasoning failure (Yu et al., 2026). Frame selection methods estimate frame relevance using motion cues, saliency scores, or query-conditioned scoring and select informative subsets for reasoning (Tang et al., 2025b; Guo et al., 2025a). Retrieval-based pipelines index video segments externally and retrieve the most relevant clips for the query (Ataallah et al., 2024; Yuan et al., 2025a).

4.5 Inference-Time Intervention

Inference-time intervention methods work during generation instead of training, which has been popular in recent years in all LLM modalities. Chain-of-thought (CoT) prompting, for example, inserts reasoning demonstrations or instructions such as "think step by step" to structure how the model produces tokens (Wei et al., 2022; Zhang et al., 2022).

For Vid-LLMs, the prompt encourages intermediate reasoning steps that reference specific video observations, e.g., events, object interactions, and temporal relations, before integrating them into a final conclusion (Zhang et al., 2025i; Jin et al., 2025b).

Beyond linear reasoning traces, agentic reasoning procedures organize inference as an iterative loop of planning, tool calling, and result updating (Luo et al., 2025a; Huang et al., 2024d; Ishibashi & Nishimura, 2024). In video understanding, as we mentioned before, hallucination often arises from failures in summarization and planning rather than perception. Agentic frameworks address this by first planning which parts of the video to inspect, then performing targeted operations such as selecting relevant frames or retrieving supporting segments (Zhang et al., 2025c; Chen et al., 2025a). The final answer is produced only after sufficient visual evidence has been collected.

5 Mitigation Techniques

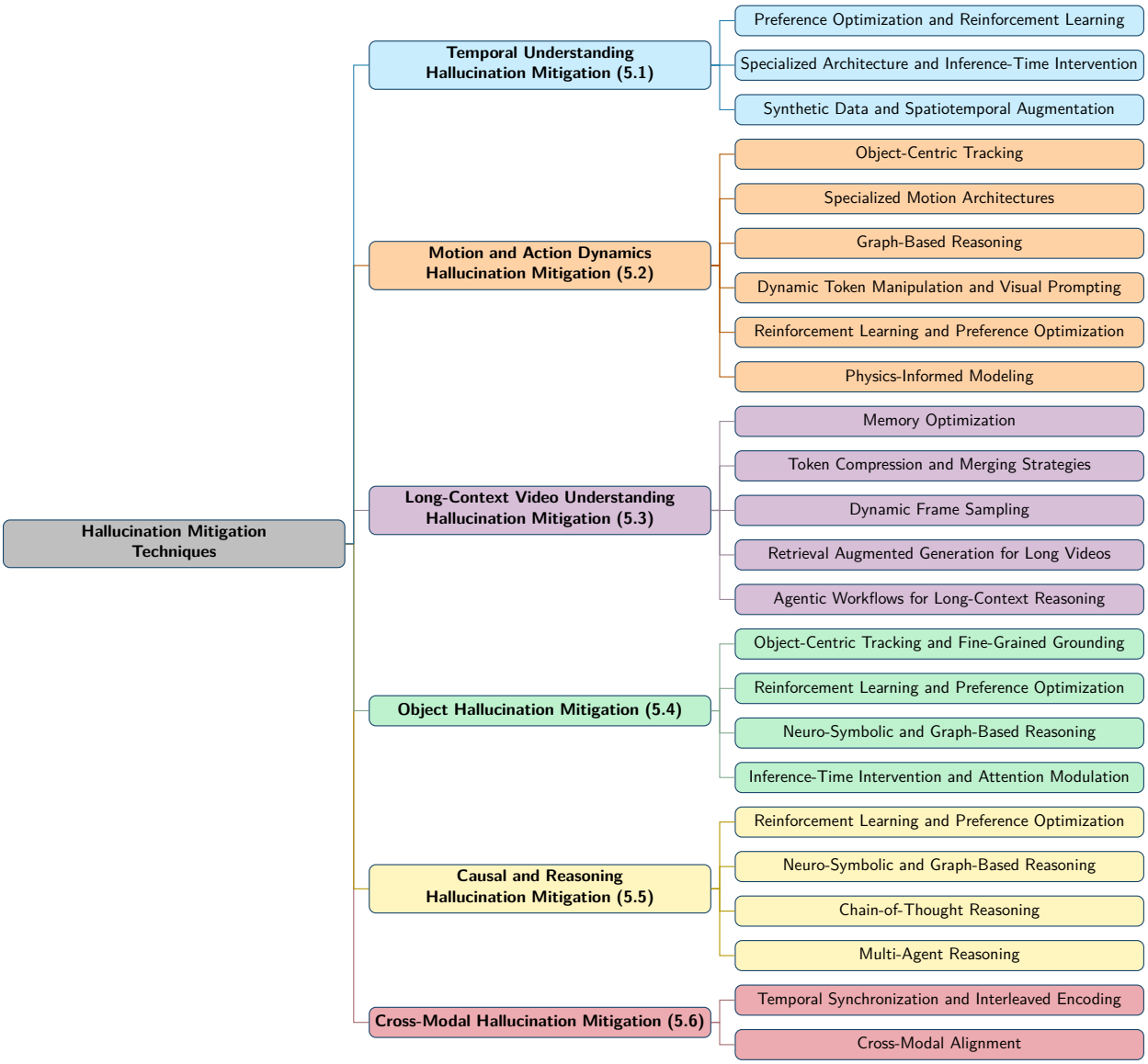

Figure 5: Taxonomy of video understanding hallucination mitigation techniques, organized by category and method.

We now review concrete mitigation techniques organized according to the hallucination categories from Section 3, examining how different methods improve model faithfulness and reliability (see Figure 5).

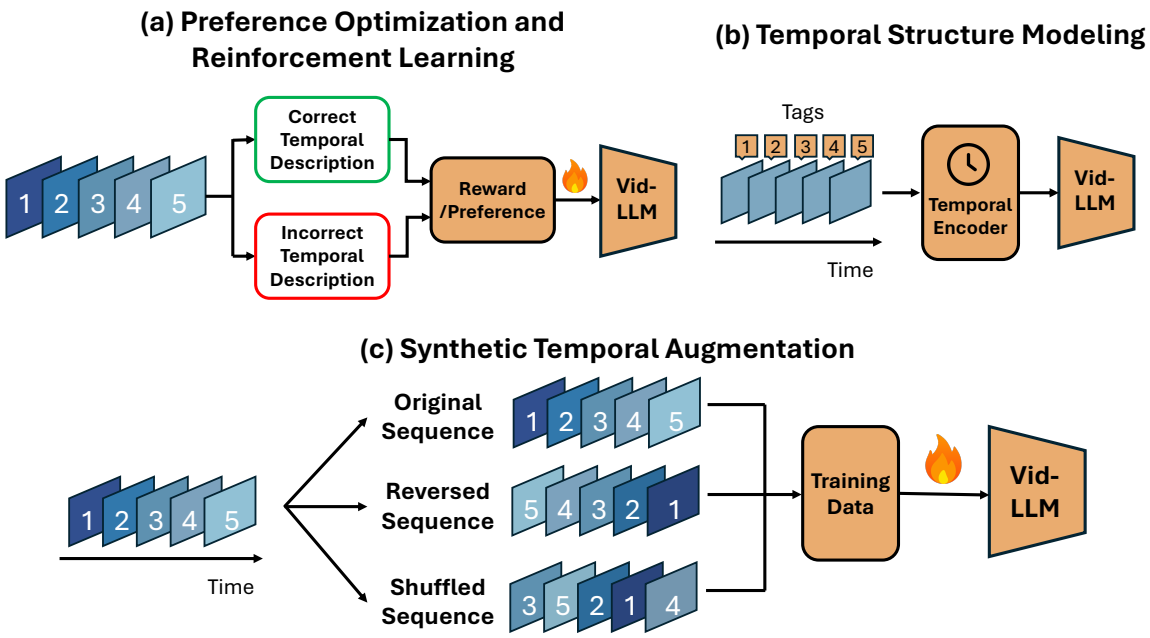

Figure 6: Illustration of representative strategies for mitigating temporal understanding hallucinations in Vid-LLMs. (a) Preference optimization and reinforcement learning methods enforce temporally grounded reasoning by rewarding correct frame ordering and penalizing temporally inconsistent predictions. (b) Temporal structure modeling introduces dedicated temporal encoders or time-aware representations that explicitly capture event ordering and timestamp information for downstream reasoning. (c) Synthetic temporal augmentation improves robustness by exposing the model to temporally perturbed training samples, such as reversed or shuffled event sequences.

5.1 Temporal Understanding Hallucination Mitigation

Temporal understanding hallucination is a major failure mode of Vid-LLMs, motivating the inventions of techniques that target temporal ordering, localization, and duration grounding. Existing approaches fall into three technical paradigms: (i) *preference optimization and reinforcement learning* with temporally grounded reward signals; (ii) *specialized architectures and inference-time interventions* that encode or diagnose temporal structure; (iii) *synthetic data generation and spatiotemporal augmentation* to expose models to temporally perturbed supervision (see Figure 6).

Preference Optimization and Reinforcement Learning Early supervised fine-tuning approaches such as TimeChat (Ren et al., 2023) and TRACE (Guo et al., 2024b) penalize false negatives under autoregressive losses, causing overfitting and poor generalization. RL-based alternatives are motivated by the insight that temporal localization quality is better captured by task-specific reward signals than by token-level cross-entropy.

Time-R1 (Wang et al., 2025k) post-trains Vid-LLMs using RL with verifiable rewards (RLVR), combining a timestamp-aware Intersection-over-Union (tIoU) signal with a reasoning-template reward optimized via Group Relative Policy Optimization (GRPO). On Charades-STA (Gao et al., 2017), Time-R1 achieves R1@0.7 of 35.3, roughly doubling TRACE and nearly tripling TimeChat. It reduces hallucination rates by 10–15% from baselines. Temporal grounding improves when the model is rewarded for evidence-consistent temporal decisions rather than matching surface-form annotations. TimeZero (Wang et al., 2025l) demonstrates that structured RL-guided reasoning can yield competitive temporal localization from as few as 2,500 training samples. However, both methods incur higher training and inference costs and struggle with ultra-long video contexts.

D^2VLM (Zeng et al., 2025b) addresses the entanglement between temporal grounding and text generation by decoupling these objectives through explicit evidence tokens that capture event-level visual semantics, forcing the model to obtain temporal evidence before generating an answer. This improves grounding accuracy over TimeChat-7B (Ren et al., 2024a) with only 1.4% computational overhead per token, though performance on episodic memory and fine-grained event boundary detection remains limited.

Subsequent work refines the RL approach along several axes. VideoChat-R1 (Li et al., 2025g) extends to multi-task reinforcement fine-tuning, jointly optimizing temporal grounding, object tracking, and question-answering, yielding gains over Video-R1 (Feng et al., 2025). MOSSChat-V (Tao et al., 2025) addresses "temporal hacking", where models bypass temporal reasoning to guess outcomes directly, through a process-reasoning reward based on subsequence Dynamic Time Warping (DTW) that supervises intermediate temporal logic. Li et al. (2025c) combine discrete semantic rewards with continuous tIoU signals alongside variance-aware data selection, surpassing Video-R1 on temporal grounding with substantially less training data.

Additional efforts explore hybrid off-policy sampling and soft advantage estimation (Li et al., 2025i), multi-task RL with task-specific localization rewards (Wu et al., 2025c), and metric-based direct preference optimization through chain-of-tasks decomposition (Lee et al., 2025). These studies confirm a shift from static supervised objectives toward adaptive, reward-driven post-training. The main strength is its flexibility: post-training alignment can layer on top of existing Vid-LLMs without architectural redesign. On the other hand, reliable rewards are very tricky to design, long-horizon reasoning increases training instability, and scalability to ultra-long contexts remains very challenging.

Specialized Architecture and Inference-Time Intervention Architectural approaches modify what the model *represents* by embedding temporal structure directly into its token vocabulary, attention mechanisms, or decoding pipeline. Their thought was, if time is only weakly encoded in the input, no amount of post hoc alignment can fully recover precise chronology.

Early work VTimeLLM (Huang et al., 2023) introduced a three-stage curriculum that progressively sharpens temporal perception through dedicated boundary detection tasks, improving temporal grounding over Video-LLaMA (Zhang et al., 2023b). Models that recognize content well may still fail to segment event boundaries, indicating that temporal hallucination is partly representational. However, VTimeLLM represents timestamps as plain text and relies on detected shot boundaries, limiting intra-shot precision.

LITA (Huang et al., 2024b) combines relative time tokens with a SlowFast architecture, compressing the token count from 25,600 to 356 while nearly doubling mIoU on ActivityNet-RTL (Huang et al., 2024c). Chrono (Rodriguez et al., 2024) shows that a simpler, model-agnostic approach—interleaving frame embeddings with absolute integer timestamps—can outperform complex specialized modules with only 19M trainable parameters via LoRA. These results suggest that some temporal hallucinations arise not from weak visual features but from the model's inability to interpret temporal metadata. Grounded-VideoLLM (Wang et al., 2024a) extends this with a dual-stream architecture that encodes spatial and temporal components through discrete temporal tokens sharing an embedding space with the language model.

VTG-LLM (Guo et al., 2024a; 2025c) tackles concept shifts from identical output targets for varying inputs by integrating sequence-time embeddings, absolute-time tokens, and slot-based compression enabling denser frame sampling without expanding the context window. GeLM (Chen et al., 2024d) augments the vocabulary with dynamic grounding token pairs fused with visual features through a dual-branch saliency-similarity module, more than doubling IoU@0.3 on MULTIHOP-EGOQA (Chen et al., 2025c) relative to TimeChat (Ren et al., 2024a), which is important for settings where the answer depends on several disjoint moments.

Two recent directions go beyond discrete tokens. VTune (Jung et al., 2024) reformulates temporal grounding as a verification process requiring models to confirm or correct query-moment alignments at inference time, substantially improving TimeChat's consistency on Charades-CON (Jung et al., 2025)—revealing that specialized timestamp encoding alone does not guarantee robust temporal comprehension. DisTime (Zeng et al., 2025c) models time as continuous probability distributions rather than discrete tokens: its distribution-based decoder handles boundary ambiguity while adding less than 1% additional parameters, surpassing discrete-

quantization methods on Charades-STA (Gao et al., 2017). Continuous decoding is better suited to uncertain or gradually evolving events, though it remains constrained by frame-level evidence accuracy.

Complementary refinements include iterative timestamp refinement with auxiliary regression losses (Wang et al., 2024j). The field traces a trajectory from text-based timestamps through discrete tokens and dual-stream architectures to continuous distributional modeling. Architectural methods offer stronger temporal inductive bias than post-training alignment, but require more invasive modifications or denser visual processing. A promising direction is unifying these encoding paradigms for fine-grained boundary localization, multi-hop reasoning, and verification-based consistency without task-specific modifications.

Synthetic Data and Spatiotemporal Augmentation Data-centric approaches tackle temporal hallucinations at their source: the training distribution. These methods create supervision that foregrounds order, progression, and temporal coherence through synthetic or augmented videos with controlled temporal variations.

Sparrow (Yin et al., 2024) augments real videos with synthetic text-to-image sequences—rendering instructions as image sequences mimicking video frames—reducing required training samples by 85% while improving temporal reasoning on Video-MME over Video-ChatGPT (Maaz et al., 2023). VISTA (Ren et al., 2024b) and COSA (Chen et al., 2023b) address models trained on short clips that fail on long videos by constructing pseudo long-form sequences from short samples, strengthening event-level alignment without large native long-video corpora.

TimeWarp (Vani et al., 2025) generates preference pairs by shuffling and reversing event sequences within video narratives, extending the STIC (Deng et al., 2024) self-training framework to video. This reduces hallucination rates by 7.5% on temporal benchmarks at a cost of approximately 60 A100 GPU hours per model. It is particularly effective against shortcut learning, where a model answers correctly from scene cues while ignoring chronology. Complementary works inject temporally sensitive tasks to reduce reliance on static visual bias (Wang et al., 2025p; Zhou et al., 2025a). Perturbation-based augmentation exposes whether the model truly understands sequence order, though aggressive perturbations can drift from realistic video dynamics.

A third branch shifts from recognition to prediction. Next-Event Prediction (NEP) (Wang et al., 2025c) trains models to generate textual summaries of future video segments conditioned on past frames, forcing internalization of event dynamics and causal progression. DIBS (Wu et al., 2024a) pursues a similar goal through diverse LLM-generated event-centric pseudo captions with online boundary refinement.

T3 (Li et al., 2024b) finds that the temporal reasoning bottleneck often resides in the language model rather than the visual encoder: by synthesizing diverse temporal tasks in pure text from image-text datasets, T3 surpasses ShareGPT4Video (Chen et al., 2024c) on TempCompass (Liu et al., 2024f) without any video data.

Data-centric approaches are more scalable than RL and less invasive than architectural redesign, but if synthetic supervision contains temporal bias or implausible event chains, it can regularize the model in the wrong direction. Careful control over temporal structure—through perturbation, composition, or modality transfer—proves more important than sheer data volume.

5.2 Motion and Action Dynamics Hallucination Mitigation

Motion and action dynamics hallucination emerges when Vid-LLMs fail to capture fine-grained temporal evolution, object trajectories, and physically grounded motion patterns. We group existing mitigation techniques into six categories (see Figure 7): (i) *object-centric tracking*; (ii) *specialized motion architectures*; (iii) *graph-based reasoning*; (iv) *dynamic token manipulation and visual prompting*; (v) *reinforcement learning and preference optimization*; and (vi) *physics-informed modeling and 4D reconstruction*.

Object-Centric Tracking Object-centric tracking replaces coarse frame-level features with explicit per-entity trajectory representations, since action errors often stem from the model not knowing who moved, where, and in what order. PiTe (Liu et al., 2024d) tracks three key points per object across frames and projects these trajectories into the language embedding space via a dedicated trajectory projector. Sup-

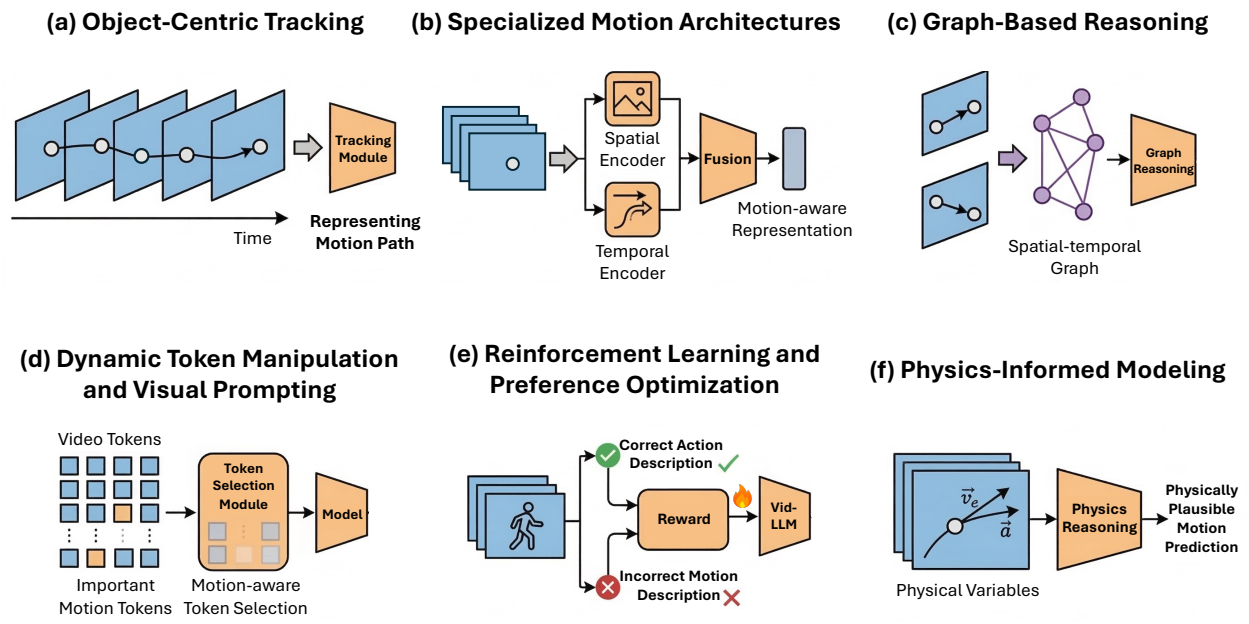

Figure 7: Illustration of representative methods for mitigating motion and action dynamics hallucinations in Vid-LLMs. Each panel depicts a representative strategy corresponding to one category discussed in this subsection, including object-centric tracking, specialized motion architectures, graph-based reasoning, dynamic token manipulation and visual prompting, reinforcement learning and preference optimization, and physics-informed modeling.

ported by an automatic annotation pipeline yielding the PiTe-143k dataset and k-means++ clustering to manage computational overhead, PiTe demonstrates superior zero-shot performance on video question answering, temporal grounding, and dense captioning compared to static-alignment baselines such as Video-LLaMA (Zhang et al., 2023b). However, its reliance on sparse keypoint tracking limits performance with small or heavily occluded objects.

VideoOrion (Feng et al., 2024) extends to full object-centric tokenization through a two-branch design: a video-centric branch captures global context while an object-centric branch leverages a detect-segment-track pipeline (GroundingDINO, SAM, XMem) to extract disentangled object tokens. This substantially improves performance across MVBench (Li et al., 2024a), EgoSchema (Mangalam et al., 2023), Perception-Test (Patraucean et al., 2023), VideoMME (Fu et al., 2025), and ActivityNet-QA (Yu et al., 2019) over spatial-temporal pooling baselines, at a roughly 38% increase in computation time.

iMOVE (Li et al., 2025e) introduces the first motion-aware instruction-tuning dataset (iMOVE-IT) with fine-grained instance spatiotemporal annotations, paired with Event-aware Spatiotemporal Efficient Modeling and Relative Spatiotemporal Position Tokens. This reduces visual tokens by over 34% via adaptive pooling while preserving motion fidelity, and unifies spatial grounding, temporal grounding, and dynamic captioning through mutual-supervision tasks. However, fixed event segmentation granularity limits applicability to ultra-long videos. The key lesson: object grounding is most effective when architecture and training data are jointly designed around motion.

Tracklet-phrase contrastive alignment (Chang et al., 2025a) improves object and action F1 on MiraData-9k (Ju et al., 2024) over Vista-LLaMA (Ma et al., 2023). Complementary works include trajectory-guided attention for detecting physical implausibility (Motamed et al., 2025) and slot-based tokenization of object dynamics (Xu et al., 2024a). MotionBench (Hong et al., 2025a) exposes remaining limitations on fine-grained motion tasks. These approaches force temporal reasoning to be indexed by entities instead of latent global tokens; their main limitation is scalability in crowded scenes, under heavy occlusion, or with amorphous motion regions.

Specialized Motion Architectures Whereas object-centric tracking augments the input representation, specialized motion architectures modify the model's internal processing pathways to better capture high-frequency inter-frame dynamics. Rasekh et al. (2025) observed that joint spatiotemporal attention, as in InternVideo2-Chat (Wang et al., 2024k), is insufficient for distinguishing temporally mirrored actions. Their STAVEQ2 integrates stacked temporal attention (STA) modules within the vision encoder, processing patch embeddings across frames before tokens reach the language model, achieving an 11-point accuracy gain on SSv2-T10 with up to four times fewer attention heads. Early temporal processing is critical when mirrored or near-identical actions must be distinguished, though performance degrades on extremely long videos.

MASH-VLM (Bae et al., 2025) instead targets entangled causal attention within the language model: DST-attention restricts direct interactions between spatial and temporal tokens, while Harmonic-RoPE harmonizes positional IDs across multimodal tokens. This reduces hallucination rates by 12–17% on UNSCENE (Bae et al., 2025) but remains constrained by fixed token budgets beyond 100 frames.

Complementary refinements include: a GOP encoder jointly fusing RGB and motion features within group-of-pictures units (Zhao et al., 2025c), surpassing Video-LaVIT's (Jin et al., 2024) concatenation-based fusion; a diffusion-powered continuous video tokenizer preserving spatial-temporal dynamics (Ge et al., 2024); an appearance-free optical flow stream with action-centric contrastive learning (Chen et al., 2023a), improving NExT-QA (Xiao et al., 2021) accuracy from 55% to 58%; and disentangled frame-level and motion-level features through noun-verb grounding (Liu et al., 2024b).

At the systems level, Li et al. (2025h) enable 16 FPS processing via a high-frame-rate aligner with block matrix decomposition, while Hu et al. (2025b) improve fine-grained sensitivity through self-supervised fragment-level tasks. The central trade-off is computational: motion-specialized encoders improve fidelity but increase memory use, reduce compatibility with generic backbones, or degrade on very long videos.

Graph-Based Reasoning Graph-based methods construct explicit structured representations, such as scene graphs where nodes encode objects and edges capture spatial or temporal relationships, to enable reasoning about fine-grained interactions rather than implicit holistic feature matching. Finsta (Fei et al., 2024b) represents both texts and videos as fine-grained scene graph structures unified into a holistic graph, with a spatial-temporal Gaussian differential Graph Transformer that differentiates between moving and stationary objects (Sun et al., 2022). This improves long-form video QA accuracy by 7–9% over the base Vid-LLM, but its performance degrades with noisy annotations and videos containing more than eight sequential actions.

GHR-VQA (Brilli et al., 2025) constructs human-rooted video-level graphs linking frame-level scene graphs via human nodes to a global root, with a hierarchical Conditional Relational Network and HetEdgeGAT encoder for cross-frame message passing. This achieves substantial gains on AGQA (Grunde-McLaughlin et al., 2021) over prior graph-based approaches (Khan et al., 2023), since many motion questions are implicitly actor-centric. However, the single-root constraint limits scalability to multi-actor scenes.

Video-STR (Wang et al., 2025i) bridges graph-based reasoning with RL by constructing inter-object relation graphs and training via GRPO with task-specific rewards. Its rotation-invariant graph representation addresses the lack of explicit spatial rewards in Video-R1 (Feng et al., 2025), though graph construction introduces computational overhead.

Further extensions in this category include: a dynamic graph transformer modeling object-level temporal dynamics (Xiao et al., 2022) beyond the static graphs of HQGA (Peng et al., 2022); temporally coherent event graphs preserving trajectory and velocity attributes (Masala & Leordeanu, 2025); multi-scale temporal modeling with visual-action semantic-aware modules (Liu et al., 2025a); relation-aware graph attention mechanisms (Wang et al., 2023c); and motion-focused video-language learning with verb-variation strategies (Doughty et al., 2024). Graph-based approaches excel at compositional questions involving contact, order, or multi-object dependencies, but noisy detections and annotation errors can themselves become a source of hallucination. Hence, graph reasoning is most useful when paired with stronger perceptual front ends.

Dynamic Token Manipulation and Visual Prompting A complementary line of work manipulates visual token representations or augments input with explicit visual prompts that direct attention toward active motion regions. DynImg (Bao et al., 2025) composes key frames into a single dynamic image while using non-key frames as pixel-level temporal prompts highlighting rapidly changing regions, with a 4D rotary positional embedding preserving spatiotemporal order. Unlike Video-ChatGPT (Maaz et al., 2023) and Video-LLaMA (Zhang et al., 2023b), DynImg enables fine-grained pixel-level interactions before feature extraction, achieving substantial gains on MVBench motion subtasks. It performance starts to degrade when more than four non-key frames are used.

MotionSight (Du et al., 2025a) decouples object motion from camera motion at zero-shot inference time, since models often hallucinate object movement when the real cause is so-called ego-motion, i.e., the motion of a camera relative to its environment. This yields a 14.3% gain on camera motion tasks in MotionBench (Hong et al., 2025a) with modest additional inference latency. ST-LLM (Liu et al., 2024c) shows that LLMs can directly model spatial-temporal token sequences without additional temporal modules, using dynamic masking for masked video modeling that improves MVBench (Li et al., 2024a) accuracy over module-heavy approaches (Luo et al., 2023). Domain-specific heuristic prompts (Yu et al., 2024b) offer lightweight alternatives. These methods are training-free with minimally invasive front-end corrections. On the other hand, when motion salience is ambiguous, manually designed prompting can suppress useful context as easily as it highlights relevant change.

Reinforcement Learning and Preference Optimization RL and preference optimization penalize incorrect action dynamics through motion-sensitive reward signals. Standard final-answer rewards are too coarse to penalize subtle mistakes in trajectory or event ordering. Invert4TVG (Chen et al., 2025l) augments RL training with verb completion, action recognition, and video description, which are derived from grounding annotations without external supervision. It reinforces action-semantic consistency while reducing GPU memory by 15%.

VideoCap-R1 (Meng et al., 2025) extends RL to open-ended captioning through GRPO with a structured thinking paradigm, using dual reward mechanisms that includes an LLM-free think scorer and an LLM-assisted caption scorer, to yield strong F1 improvements on CAREBENCH (Xu et al., 2024b) with only 1.5k training samples. Video-CoM (Rasheed et al., 2025) introduces a Chain of Manipulations enabling iterative visual interactions through temporal segment finding, keyframe isolation, and spatial zooming, supervised by step-level rewards for temporal IoU, frame recall, and spatial IoU. It is effective because motion understanding is often a sequential search problem.

VideoPerceiver (Zhao et al., 2025a) enhances temporal sensitivity through key-information-absent video construction and contrastive learning, then applies a comparative GRPO reward enforcing superior response quality from complete videos over temporally degraded inputs. This yields a 22.9% improvement on VRU-Accident (Kim et al., 2025) with considerable computational overhead. VDCAgent (Wang et al., 2025f) eliminates dependence on expensive external models via agentic self-reflection and curriculum DPO, while OwlCap (Zhong et al., 2025) proposes a Caption Set Equivalence Reward jointly optimizing correctness and completeness beyond motion-only (Wang et al., 2024d) or event-coverage-only rewards.

These works together represent a shift from task-specific binary rewards to rewards that capture subtle physical plausibility constraints (trajectory consistency, velocity coherence, biomechanical feasibility, etc.) together with external annotators at affordable costs.

Physics-Informed Modeling Physics-informed modeling directly embeds geometric and physical priors into the reasoning pipeline, constraining the hypothesis space so that implausible motion interpretations are rejected. Zhou et al. (2025c) address Vid-LLMs' inability to perform dynamic spatial reasoning by introducing the DSR Suite framework with an automated training data pipeline and a Geometry Selection Module employing two stacked Q-Formers. The module condenses question semantics and selectively extracts task-relevant geometric knowledge from 4D reconstruction priors, introducing only 32 geometry tokens to balance information density and computational cost. MASS (Wu et al., 2025d) extends this with depth-based 3D encoding, visual grounding, and an explicit motion tracker. PhyVLLM (Zhan et al., 2025a)

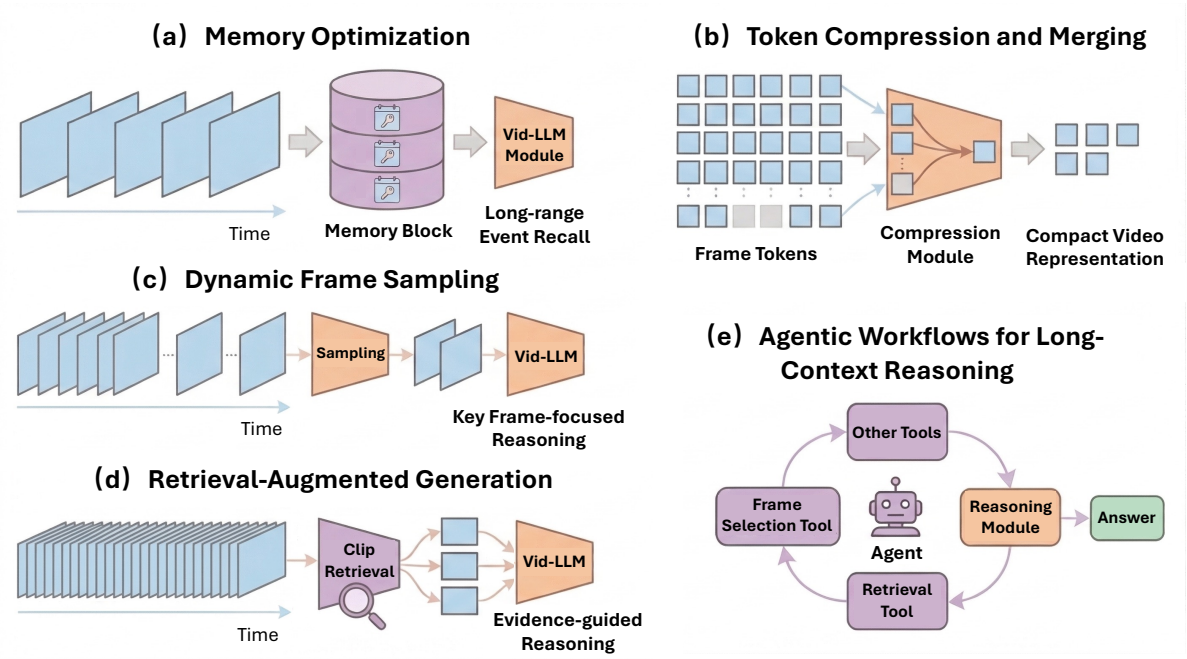

Figure 8: Illustration of representative methods for mitigating long-context video understanding hallucinations in Vid-LLMs. Each panel depicts a representative strategy corresponding to one category discussed in this subsection, including memory optimization, token compression and merging, dynamic frame sampling, retrieval-augmented generation, and agentic workflows for long-context reasoning.

further extends them with continuous-time modeling through motion-appearance disentanglement, Neural ODE-based dynamics, and self-supervised physical consistency losses.

The difficulty lies upstream: physics-informed methods depend on the quality of depth estimation, reconstruction, or motion decomposition. When these priors are noisy, the model will inherit more errors than flexible learned features would. The challenge is making integration uncertainty-aware and robust to real-world sensing noise.

Summary. These reviewed mitigations differ in where they intervene: object-centric tracking and specialized architectures strengthen perception; graph-based reasoning and physics-informed modeling impose structure on interpretation; dynamic token manipulation offers cheaper inference-time correction; and RL reshapes training objectives toward motion-sensitive fidelity. Tracking and graph pipelines improve grounding but add preprocessing cost; specialized architectures improve accuracy but reduce efficiency; prompting is lightweight but heuristic; and RL is certainly powerful but expensive. A promising direction is, perhaps, a hybrid design: tracked or prompted motion cues feeding structured reasoning, with lightweight preference signals and uncertainty-aware physical priors.

5.3 Long-Context Video Understanding Hallucination Mitigation

As we've mentioned before, long-context video understanding poses a fundamental challenge because exhaustive frame processing is computationally infeasible, so models rely on aggressive sampling, compression, or memory mechanisms. Hallucinations arise from failures to retrieve, retain, or attend to temporally relevant evidence. We break corresponding mitigation techniques into five subcategories: (i) *memory optimization*; (ii) *token compression and merging*; (iii) *dynamic frame sampling*; (iv) *retrieval augmented generation*; and (v) *agentic workflows for long-context reasoning*.

Memory Optimization Early work by Song et al. (2023) introduced a training-free framework combining a memory buffer with merge-based consolidation, but non-learnable frame merging caused over-smoothing

in long videos. VideoLLaMB (Wang et al., 2024o) addresses this with recurrent memory bridges, which are learnable memory tokens within bridge layers, alongside a SceneTiling algorithm for semantic video segmentation, ensuring memories respect scene structure rather than arbitrary chunk boundaries. On NIAVH, VideoLLaMB substantially outperformed MA-LMM (He et al., 2024). A concurrent variant (Wang et al., 2024m) refined this for streaming with linearly scaling recurrent memory caches.

Subsequent work shifted toward human-inspired and online memory mechanisms. Flash-VStream (Zhang et al., 2024a) introduced a Spatial-Temporal-Abstract-Retrieved (STAR) memory that dynamically compresses visual information online, achieving real-time inference with lower VRAM. An enhanced version (Zhang et al., 2025d) decomposed memory into Context Synopsis Memory for temporal aggregation and Detail Augmentation Memory for spatial detail, enabling sub-second responses. It effectively addresses hallucination that arise when coarse narrative state and local visual detail are compressed. ∞-Video (Santos et al., 2025) augments video Q-formers with continuous-time long-term memory consolidation based on a Gibbs density function, dynamically allocating higher memory granularity to relevant segments, though fine-grained temporal reasoning remains limited when sampling is coarse.

Video-SALMONN S (Sun et al., 2025c) replaces token merging with a test-time training layer using Hessian-free optimization to continually update memory representations, processing over one million tokens with fixed memory overhead at approximately 14% additional computational cost. VideoMem (Jin et al., 2025c) addresses the rigidity of static memory with adaptive memory management trained via Progressive Grouped Relative Policy Optimization (PRPO) with Temporal Cascading Rewards, achieving a 3.1$\times$ training speedup. The key insight is: long-context faithfulness depends less on stored tokens than on learning when to retain, rewrite, or discard them.

A parallel line of work targets KV cache compression for streaming. Yang et al. (2025c) introduced query-agnostic KV cache compression with fixed memory footprint, while Chen et al. (2025g) replaced uniform segmentation with dynamic semantic partitioning. Xu et al. (2025a) aligned training with streaming inference via overlapped short-chunk attention, achieving competitive performance against GPT-4o mini on infinite-stream evaluation. Chen et al. (2025i) developed MR-SP to scale RL training to hour-long videos on a single A100 node via sequence parallelism with cached video embeddings.

Additional strategies include multi-stage long-context training curricula (Xue et al., 2024), hybrid Mamba-Transformer architectures for linear-complexity processing (Ren et al., 2025), task-aware KV sparsification (Qin et al., 2025b;a), and agentic frameworks with adaptive token selection (Yang et al., 2025b). The central trade-off remains between memory footprint and retrieval fidelity: aggressive compression risks losing transient events critical for fine-grained reasoning, while richer memory representations undermine real-time applicability.

Token Compression and Merging Strategies Token compression reduces representation size by fusing or aggregating visual tokens after encoding. The difficulty is that naive pooling often destroys temporal order, object identity, and scene transitions, which really matter for faithful reasoning.

LongVLM (Weng et al., 2024) decomposes long videos into segments and applies hierarchical token merging, reducing token counts by roughly two orders of magnitude, but its fixed compression ratio limits adaptability. Dynamic-VLM (Wang et al., 2024c) adaptively allocates between 16 and 576 tokens per frame based on video length, substantially improving over LLaVA-OneVision but remaining constrained by a fixed 12K token budget. ReTaKe (Wang et al., 2024g) jointly reduces temporal and knowledge redundancies through DPSelect for keyframe selection and PivotKV for KV cache compression. It outperformes MovieChat by 3–5% though degrading on fine-grained tasks such as Needle QA.

Video-XL (Shu et al., 2024) leverages LLM attention sparsity through Visual Summarization Tokens that adjust compression granularity based on information density. Wei & Chen (2024) adapt YaRN position embedding extension to visual tokens with progressive pooling, reducing memory by approximately 45% without retraining. BIMBA (Islam et al., 2025) is the first SSM-based token compression approach, which uses bidirectional selective-scan at $O(N)$ complexity and compressing tokens by 16 times. CrossLMM (Yan et al., 2025) decouples visual token processing from the LLM through dual cross-attention, producing text-conditioned compressed representations with sub-linear memory growth. Dahal et al. (2025) combine tem-

poral pooling with DPO, compressing video into a single visual token per second with a 23 times token reduction, though reliability depends on subtitle quality.

Liu et al. (2025e) introduced reconstructive token compression extending processing to over 8,000 frames while maintaining near-perfect Needle-in-a-Haystack accuracy. This work demonstrates that reconstructive compression preserves semantics better than pure sparsification. Jiang et al. (2025a) insert a Mamba-based temporal encoder between image encoder and LLM, showing compression can be built into the sequence processor itself. Complementary strategies include hierarchical caching for streaming (Wang et al., 2025n), thumbnail-and-sampling schemes (Qu et al., 2024), and strictly causal two-stage compression (Chen et al., 2025f). Token compression is better suited than dynamic sampling when broad temporal coverage must be preserved, but risks semantic averaging where distinct events collapse into a single token.

Dynamic Frame Sampling Dynamic frame sampling reduces information overload at the input stage by selecting or filtering frames based on semantic relevance or information density, thus preventing redundancy-induced fabrication before tokens enter the model.

The LIVE framework (Chen et al., 2024a) introduced Streaming EOS prediction for streaming dialogue, dynamically determining when to respond or remain silent. However, its single CLS token per frame limits spatial understanding. VideoLLM-MoD (Wu et al., 2024b) introduces a Mixture-of-Depths strategy that selects the top-k most informative vision tokens per frame while routing others through residual connections, reducing FLOPs by 42% and supporting 1.7× longer context windows. Sun et al. (2024) combine CLIP-score-guided frame sampling with question-injected visual features, achieving a hallucination rate reduced by over 24% relative to MovieChat (Song et al., 2023). It comes from the key insight that relevance should be defined with respect to the question, not only video salience. SEAL (Wang et al., 2024e) decomposes long videos into compact scene, object, and action tokens with an Attention Learning module maximizing query relevance and token diversity.

Sun et al. (2025b) extended keyframe extraction to temporally coherent key clips with adaptive resolution, outperforming frame-level baselines on Video-MME and MLVU (Zhou et al., 2025b). Yao et al. (2025) reformulated keyframe selection as clip-to-frame prediction preserving scene continuity. Xu et al. (2025b) leveraged event-camera density for coarse-to-fine redundancy removal at lower cost. Chasmai et al. (2025) introduced moment sampling through a pre-trained moment retrieval model, yielding consistent gains across four long-form VideoQA benchmarks. Ok & Lee (2025) showed that vision encoder confidence scores do not correlate with downstream performance and that up to 50% of frames can be removed without significant degradation, motivating segmentation-area-based selection as a more robust alternative.

The design space continues to grow: hierarchical visual retrieval (Gao et al., 2025a), RL-based temporal sampling (Tang et al., 2025a), DPP-driven diversity-aware selection (Wang et al., 2025g), adaptive frame importance prediction (Li et al., 2025b), differentiable selection policies (Buch et al., 2025), iterative frame spotlighting (He et al., 2025b), and training-free query-aware sampling (Liu et al., 2025d). Dynamic sampling is most effective when redundancy is high and target evidence is localized; it is less effective when essential information is temporally dispersed or only identifiable after deeper reasoning.

Retrieval-Augmented Generation for Long Videos Retrieval-augmented generation (RAG) treats the video as a searchable knowledge base, extracting only query-relevant segments. This approach is especially effective when only a small fraction of a long video is relevant.

Naive RAG strategies segment videos mechanically into short clips and retrieve them independently, but they disrupt temporal continuity. Vgent (Shen et al., 2025) addresses this with a graph-based framework modeling clips as nodes interconnected by shared entities, outperforming naive RAG by 8.6% on MLVU (Zhou et al., 2025b). MemVid (Yuan et al., 2025a) maintains a key-value cache for holistic video memorization with generative context-aware query expansion operating directly on video content, reducing contradictions by 8.7% with lower GPU memory than Video-XL (Shu et al., 2025). These methods employ retrievals that apply uniform effort regardless of query complexity.

AdaVideoRAG (Xue et al., 2025a) classifies query intent and dynamically routes questions through hierarchical retrieval paths, from direct inference for simple queries to graph-based retrieval for complex multi-hop

reasoning. SceneRAG (Zeng et al., 2025a) segments videos into narrative-consistent scenes using ASR transcripts and temporal metadata for dynamic knowledge graph construction, since retrieval quality depends on whether indexed units correspond to meaningful event boundaries. Clearly, reliance on ASR quality limits its effectiveness on videos with no or sparse dialogue.

Ranasinghe et al. (2024) show that explicit entities serve as better indexing anchors than redundant captions, outperforming LLoVi (Zhang et al., 2023a). Goulas et al. (2024) replace LLoVi's joint caption processing with a linear-complexity frame-wise approach. Ma et al. (2024) reformulate long video understanding as document retrieval with multi-stage agent interaction. Luo et al. (2024) propose a training-free framework replacing redundant visual tokens with retrieved auxiliary texts (OCR, ASR, object detection), demonstrating that auxiliary text retrieval can be more efficient than additional visual tokens (Fan et al., 2024). Additional efforts decompose long videos into short-term captions for LLM-based reasoning (Ataallah et al., 2024). Retrieval-based methods shift complexity from sequence length to indexing quality, but retrieval errors are often unrecoverable, so recent systems increasingly combine retrieval with memory, graph structure, or adaptive routing.

Agentic Workflows for Long-Context Reasoning Agentic workflows turn video understanding into a multi-step process of planning, searching, observing, verifying, and revising.

Zhang et al. (2025g) framed long-video understanding as autonomous tool-using search over global, clip-level, and frame-level retrieval. It is a shifte from passive ingestion to active evidence seeking. Wang et al. (2025r) replaced caption-heavy preprocessing with direct pixel-level evidence extraction through a plan-observe-reflect loop, reducing token usage.

Several methods strengthen the agentic loop: Li et al. (2025d) leverage multi-granular clue exploration with verification-enhanced reflection; Dong et al. (2025b) let the model identify reasoning gaps and request targeted visual evidence; VideoLucy (Zuo et al., 2025) adds hierarchical memory and backtracking; VideoARM (Yin et al., 2025) integrates observe-think-act-memorize cycles with hierarchical multimodal memory; and GCAgent (Yeo et al., 2025) constructs schematic and episodic memory before answering to prevent event merging.

The literature also has works employing collaborative or uncertainty-aware reasoning. LVAgent (Chen et al., 2025a) distributes tasks across multiple specialized Vid-LLM agents. VideoAgent2 (Zhi et al., 2025) leverages uncertainty-aware CoT to trigger plan adjustment without heavy external pipelines. TimeSearch (Pan et al., 2025) learns adaptive temporal exploration rather than relying on hand-designed policies, while MR. Video (Pang & Wang, 2025) addresses multi-round exploration bottlenecks through a MapReduce-style design analyzing short clips in parallel.

Agentic workflows are especically useful when evidence is sparse or requires iterative disambiguation, while they also introduce latency, planning instability, and possible error propagation. Many successful agentic systems depend on the same components discussed above: retrieval to narrow search, memory to preserve discoveries, and compression to keep iterative inference tractable.

Summary. Long-context hallucination mitigation has evolved from static compression into adaptive systems deciding what to store, process, retrieve, or revisit. Memory methods excel when continuity matters but require robust retention policies. Sampling and compression improve efficiency but risk removing decisive evidence. Retrieval-augmented methods scale well but remain affected by retrieval misses. Agentic workflows offer the greatest flexibility and quality at the cost of latency, cost, and complexity.

5.4 Object Hallucination Mitigation

Object hallucination arises from failures to consistently aggregate object evidence across time, due to motion, occlusion, and appearance variation, which make object perception more complex than in static images. We break the solutions into four subcategories: (i) *object-centric tracking and fine-grained grounding*; (ii) *reinforcement learning and preference optimization*; (iii) *neuro-symbolic and graph-based reasoning*; and (iv) *inference-time intervention and attention modulation*.

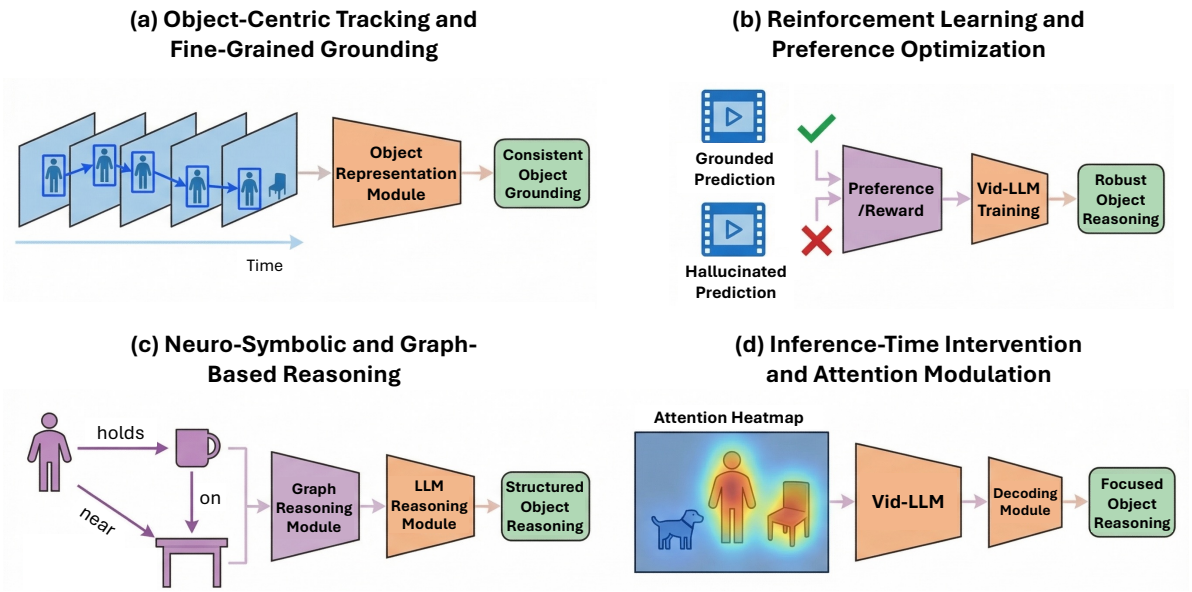

Figure 9: Overview of representative strategies for mitigating object hallucinations in Vid-LLMs. The panels illustrate four major solution paradigms discussed in this subsection: (a) object-centric tracking and fine-grained grounding for maintaining consistent object representations; (b) reinforcement learning and preference optimization that penalize unsupported object predictions during training; (c) neuro-symbolic and graph-based reasoning that structures entities and relations for grounded reasoning; and (d) inference-time intervention and attention modulation that dynamically steers the model toward visually grounded evidence.

Object-Centric Tracking and Fine-Grained Grounding The most direct remedy is to force models to explicitly localize, track, and reason about specific entities rather than relying on holistic scene representations. VideoGLaMM (Munasinghe et al., 2024) introduces a dual vision encoder (spatial and temporal) with a spatiotemporal decoder for precise mask generation, substantially improving interrogative mIoU on VidSTG (Zhang et al., 2020) over PG-Video-LLaVA (Munasinghe et al., 2023). Mask supervision forces the model to maintain correspondence between words and localized visual evidence. On the other hand, producing pixel-level masks is expensive, and its consistency struggles with objects of varying granularity across long sequences.

ObjectMLLM (Tang et al., 2025e) encodes bounding boxes and labels as quantized text tokens, nearly doubling accuracy on CLEVRER-MC (Yi et al., 2019) when augmenting VideoLLaMA 2 (Cheng et al., 2024a). RGA3 (Wang et al., 2025a) supports arbitrary visual prompts, including masks, bounding boxes, arrows, points, and scribbles, as both input and output, with a Spatial-Temporal Overlay Module propagating prompts across frames, reducing hallucination rates by approximately 10% over VideoRefer (Yuan et al., 2024). Many hallucinations arise not from failure to detect an object once but from failure to maintain its identity over time.

Gu et al. (2025) introduced text-guided temporal sampling and attribute-aware spatial activation, improving grounding on HCSTVG-v1 (Tang et al., 2021) over TubeDETR (Wasim et al., 2023). Gao et al. (2025c) coupled an MLLM with an open-vocabulary detector through a reference-semantic token, boosting m_vIoU from 8.2 to 36.2 on HC-STVG v1. Kazakos et al. (2025a) extended grounded captioning to video with spatio-temporal adapters and a temporal objectness head, yielding 20 CIDEr points on iGround (Kazakos et al., 2025b). It demonstrates that object-faithful generation benefits from temporal consistency constraints, not only spatial localization.

Complementary works include region-level token marks (Heo et al., 2025), semantically decoupled slots (Xu et al., 2024a), grid-based local-global transcription (Chowdhury et al., 2025), intention-oriented box adapters (Qiu et al., 2025), open-vocabulary detection for validation (Montes & Lozano, 2025), tokenized ob-

ject dynamics (Feng et al., 2024), and fused multi-encoder architectures (Chung et al., 2025). The consistent principle is that anchoring outputs to verifiable spatiotemporal evidence suppresses perceptual hallucinations. On the other hand, fine-grained grounding often depends on external detectors, dense supervision, or expensive token budgets, which are costly for long, cluttered videos.

Reinforcement Learning and Preference Optimization As discussed, preference optimization reshapes the output distribution by penalizing hallucinated entities, attributes, and spatial relations. VideoSAVi (Kulkarni & Fazli, 2024) pioneered self-aligned Vid-LLM training: a self-critiquing mechanism identifies reasoning errors and generates on-policy preference pairs for DPO without external supervision, yielding gains on MVBench (Li et al., 2024a) and EgoSchema (Mangalam et al., 2023). The model is trained to distinguish visually faithful descriptions from unsupported ones. Its iterative self-play is relatively computationally expensive.

VistaDPO (Huang et al., 2025a) introduced hierarchical spatial-temporal alignment at three levels: instance semantics, temporal events, and fine-grained spatial-object grounding. It applies token-level optimization via sequential KL divergence, outperforming coarse baselines on VideoHallucer (Wang et al., 2024n) and EventHallusion (Zhang et al., 2024b). PaMi-VDPO (Ding et al., 2025) eliminates offline data dependence by dynamically generating rejected samples through video augmentations, achieving competitive performance with only 10k samples. VideoPASTA (Kulkarni & Fazli, 2025) constructs targeted adversarial preference pairs contrasting dense and sparse temporal sampling, demonstrating that careful augmentation can rival methods requiring far larger preference-labeled datasets.

Oracle-RLAIF (Shi et al., 2025) replaces trained scalar reward models with an Oracle ranker providing ordinal feedback via $\text{GRPO}_{\text{rank}}$. It reports outperforming prior RLAIF approaches (Ahn et al., 2024). ViSS-R1 (Fang et al., 2025) shows that self-supervised pretext-task rewards can complement preference-based objectives without architectural modifications. Additional works include QA-driven reward signals (Liu & Wan, 2024), self-consistency-based preference pair construction (Dang et al., 2025b), and detection-and-tracking pipelines with RL policies to minimize false positive object mentions (Shankar & Surendran, 2025).

Preference-based methods are architecture-agnostic. They shape outputs indirectly: if the underlying representation lacks reliable object evidence, optimization may reduce blatant hallucinations but fail on rare objects, heavy occlusion, or fine-grained spatial distinctions. These methods need to be paired with representations that already expose object structure.

Neuro-Symbolic and Graph-Based Reasoning Neuro-symbolic and graph-based methods impose structural constraints by converting unstructured video into explicit relational graphs enforcing verifiable reasoning over entities. ReGR (Wang et al., 2023c) introduces a dual-graph architecture: a Spatiotemporal Graph Attention Network captures object dynamics, while a Multi-Relation Graph Attention Network encodes fine-grained interactions, with question-guided filtering to suppress irrelevant features. This improved accuracy on TGIF-QA (Jang et al., 2017) FrameQA but degraded on longer videos requiring complex temporal reasoning.

LGQAVE (Rongali et al., 2024) constructs dynamic spatial object graphs conditioned on masked question embeddings via a Question-aware Dynamic Graph Transformer, leveraging MiniGPT-4 for visual grounding. NExT-QA (Xiao et al., 2021) accuracy rose from 53.68% to 66.69%, and MSRVTT-QA (Xu et al., 2016) from 39.70% to 63.43%, surpassing CoVGT (Xiao et al., 2023b), which lacks question-specific conditioning. The shift from static to question-conditioned graphs is important since some object hallucinations are due to selecting the wrong subset of entities for the question at hand. However, computational overhead (289 GFlops) and reliance on MiniGPT-4 limit scalability beyond 600 frames.

Dai et al. (2025) integrated an explicit scene graph predictor, improving MOMA-QA accuracy from 76.62% to 79.66% over SeViLA (Yu et al., 2023). Complementary efforts encode textual scene graph sequences (Linok et al., 2025), leverage dual and heterogeneous graphs (Liang & Sun, 2025), and inject scene graphs into Vid-LLM prompts. What is central in this line of work is that hallucination is reduced by making object claims pass through a structured reasoning bottleneck.

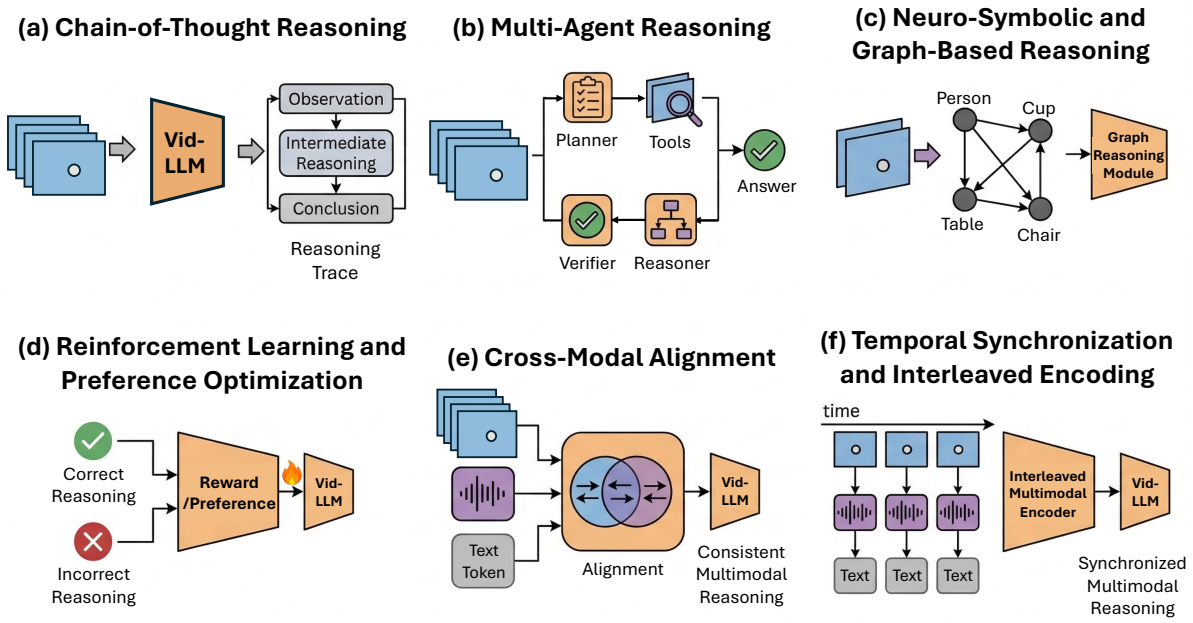

Figure 10: Representative mechanisms for mitigating causal and reasoning, and cross-modal hallucinations in Vid-LLMs. Panels (a)–(d) illustrate reasoning-level mitigation strategies discussed in Section 5.5, including chain-of-thought reasoning, multi-agent reasoning, neuro-symbolic and graph-based reasoning, and reinforcement learning with preference optimization. Panels (e) and (f) present representative approaches for cross-modal hallucination mitigation described in Section 5.6, including cross-modal alignment and temporal synchronization with interleaved multimodal encoding.

Inference-Time Intervention and Attention Modulation Inference-time interventions redirect attention toward task-relevant visual evidence during generation, offering a training-free alternative. SFA (Scan, Focus, and Amplify) (He et al., 2025a) integrates an external Video Text Spotting model with a three-stage pipeline: candidate text region detection, adaptive windowing with multimodal scoring, and region enhancement. SFA achieved up to 45.80% improvement over prior methods, including a 10.85% accuracy gain over Qwen2.5-VL-7B on RoadTextVQA (Tom et al., 2023), surpassing GAT (Zhang et al., 2025h). Instead of trusting learned attention maps, it explicitly re-centers decoding on localized visual-textual content, though effectiveness is bounded by the text spotting model's accuracy.

Jung et al. (2023) augmented visual frame representations with externally retrieved textual descriptions, reducing hallucinated object mentions in dense captioning. Inference-time interventions offer zero training cost, modularity, and compatibility with arbitrary base models, but remain narrowly scoped. Their success hinges on reliably identifying the right evidence.

Summary. Now let us compare the subcategories. Object-centric grounding is very direct and faithful, also realatively expensive. Preference optimization is more modular but cannot compensate for weak visual representations. Graph-based reasoning improves interpretability but relies on accurate scene parsing. Inference-time intervention is training-free and relatively narrow-scoped.

5.5 Causal and Reasoning Hallucination Mitigation

Causal and reasoning hallucination mitigation addresses failures where models generate unsupported causal relations, intentions, or abstract conclusions. We subcategorize its mitigation techniques as (i) *reinforcement learning and preference optimization*; (ii) *neuro-symbolic and graph-based reasoning*; (iii) *chain-of-thought reasoning*; and (iv) *multi-agent reasoning frameworks*.

Reinforcement Learning and Preference Optimization RL and preference optimization penalize logically inconsistent or visually ungrounded reasoning traces during post-training. Video-R1 (Feng et al., 2025) established the viability of RLVR for video reasoning but also exposed three challenges: holistic captions erasing temporal structure, training data solvable from textual priors alone, and CoT traces grounded in commonsense rather than visual evidence. Subsequent work has systematically addressed these limitation.

VideoRFT (Wang et al., 2025e) introduces a cross-modal CoT refinement stage where an MLLM revises initial reasoning by re-examining the video, with a semantic-consistency reward enforcing visual-textual alignment, yielding consistent gains over Video-R1 (Feng et al., 2025). Video-R2 (Maaz et al., 2025) adds a Temporal Alignment Reward and consistency gating that link reasoning trace claims to timestamps and the final answer, improving think-answer consistency across eleven benchmarks. Addressing *visual thinking drift*, Luo et al. (2025b) proposed the Visual Evidence Reward where an auxiliary judge evaluates whether intermediate steps reference verifiable visual evidence, achieving up to 9.0% gains on TempCompass.

Standard binary rewards cannot distinguish partially correct reasoning from fabricated chains. TW-GRPO (Dang et al., 2025a) enables finer-grained credit assignment through token-level importance weighting based on information entropy with multi-level soft rewards, achieving 18.8% improvement on CLEVRER (Yi et al., 2019) over Video-R1. VerIPO (Li et al., 2025j) employs a Rollout-Aware Verifier within an iterative GRPO-Verifier-DPO loop, making DPO training 7× faster than GRPO. Reflection-V (Jian et al., 2025) tackles visual attention decay in long reasoning chains through cold-start initialization with visual reflection data and a visual attention-based reward.

ReWatch-R1 (Zhang et al., 2025a) pairs multi-agent data synthesis simulating re-watching behavior with an Observation and Reasoning reward, reducing text-only solvability from 68.9% to 29.4%. VidBridge-R1 (Chen et al., 2025e) identifies a conflict between convergent QA and divergent captioning objectives, introducing intermediate proxy tasks (e.g., masked event inference and segment isolation) while eliminating the supervised fine-tuning stage. Conan (Ouyang et al., 2025b) combines multi-scale frame categorization with difficulty-aware progressive training. It reports surpassing Video-R1 by 10%+ across six multi-step reasoning benchmarks.

ReAgent-V (Zhou et al., 2025d) shifts optimization to inference time through a modular multi-agent framework generating real-time rewards via Entropy-Calibrated Relevance Scores, achieving improvements with only 45% of the training data. Complementary efforts include reinforced causal keyframe search (Zhou et al., 2025e), rank-based RLAIF (Shi et al., 2025), and self-training with preference optimization (Zohar et al., 2024). From these works, we can observe progression from simple binary outcome rewards toward sophisticated supervision of intermediate reasoning.

Neuro-Symbolic and Graph-Based Reasoning Neuro-symbolic and graph-based approaches address causal hallucinations by constructing explicit structured representations that ground inference in observable entities and relationships. It makes inferential chains auditable by design.

Some hallucinations are confounding problems where the model mistakes correlation for causality. To alleviate this, EIGV (Li et al., 2022d) partitions visual scenes into causal and environment components through a Structural Causal Model, enforcing equivariant and invariant grounding via contrastive learning, improving over invariant-only approaches (Li et al., 2022c) on MSVD-QA (Xu et al., 2017). VCSR (Wei et al., 2023b) applies causal front-door interventions through a Question-Guided Refiner and Causal Scene Separator, achieving gains on predictive and counterfactual tasks over methods (Liu et al., 2022a) that neglected temporal causal refinement. Lyu et al. (2023) leverage Semantic Role Labeling (SRL) to guide multi-step retrospective-prospective reasoning, dynamically attending to different SRL arguments and selecting relevant frames. This substantially improves TrafficQA (Xu et al., 2021) accuracy over prior graph-convolutional methods, though reliance on SRL parser quality limits generality.

GHR-VQA (Brilli et al., 2025) constructs human-rooted video-level graphs with hierarchical conditional relational networks, efficiently modeling human-object interactions though limited to single-actor scenes. Tang et al. (2024) extend with dynamic graph convolution transformers. Broader explorations include entailment-tree frameworks (Sanders et al., 2024), scene-graph and knowledge-graph fusion (Taluzzi et al.,

2025), scene graph-guided self-training (Qiu et al., 2024), theory-of-mind reasoning (Chen et al., 2024g), and causal-aware temporal grounding (Dong et al., 2025a).

The advantages of these methods include transparency and auditable reasoning chains, but they come at the cost of requiring clean scene decomposition, reliable detection and extraction, and accurate parsing. These methods are strongest when the reasoning problem has clear compositional structure and where interpretability is important.

Chain-of-Thought Reasoning CoT reasoning decomposes complex inferential tasks into intermediate steps, grounding each logical transition in observable evidence. Note that free-form textual CoT (which ignores visual evidence) can actually worsen hallucination here. Only when intermediate reasoning is visually grounded does the CoT reduce hallucination.

ViTCoT (Zhang et al., 2025i) uses a video-text interleaved paradigm that re-examines key frames at each reasoning step, achieving 5.5% improvement over baseline CoT. Causal questions often require returning to evidence after a hypothesis is formed. ChainReaction (Parmar et al., 2025) tackles causal-why video QA through a modular two-stage architecture: a Causal Chain Extractor generates chains grounded in Structural Causal Models, and a Causal Chain-Driven Answerer derives responses. While outperforming monolithic baselines, approximately 38% of errors originate from misinterpretation of object roles during chain extraction.

CoT-Vid (Jin et al., 2025a) introduces dynamic inference path routing determining whether a question requires multi-step reasoning or direct answering, with video self-consistency verification. This training-free framework outperforms Video-R1 by 9.3% on EgoSchema (Mangalam et al., 2023), avoiding unnecessary reasoning cost on simple queries.

Liu et al. (2025b) introduce video-grounded entailment tree reasoning. It achieves a competitive performance with 257× fewer parameters than VideoAgent (Wang et al., 2024h). Liang et al. (2025) refine imperfect MLLM-generated chains via a three-phase filter-and-learn pipeline. Wang et al. (2023a) replace visual embeddings with text descriptions and a token bottleneck, achieving 5× inference speedup. Other works include perception-to-cognition pipelines (Fei et al., 2024a), atomic task planning (Chen et al., 2025j), reasoning chain distillation (Shi et al., 2024), sub-task-oriented instruction tuning (Wang et al., 2025j), counterfactual thought chains (Zhang et al., 2025j), and decomposition-based subquestions (Kamoto et al., 2025). From these works we can observe that CoT helps only when the chain remains tethered to visual evidence.

Multi-Agent Reasoning Multi-agent frameworks extend CoT to collaborative, iterative processes where specialized agents debate, verify, and refine intermediate inferences, directly targeting error propagation where an early causal misattribution corrupts all downstream conclusions.

MoReVQA (Min et al., 2024) decomposed video QA into fixed stages that proved brittle when queries did not conform. CAViAR (Menon et al., 2025) shifts to adaptive, iterative execution: a reasoning agent dynamically composes video-processing modules based on intermediate results, and then a natural-language critic evaluates sampled reasoning traces. CAViAR achieves 77.2% on LVBench (Wang et al., 2025h) versus 48.7% for direct inference.

From different perspectives, Xie et al. (2025) orchestrate diverse prompting strategies with an MLLM-based evaluator for compositional and causal reasoning without task-specific training. ReAgent-V (Zhou et al., 2025d) generates evaluation reports during inference as dynamic rewards, employing reflection across conservative, neutral, and aggressive perspectives. VideoAgent (Fan et al., 2024) frames video reasoning as iterative collaboration where specialized agents collaboratively select, analyze, and refine reasoning through reward-driven feedback.

Once again, multi-agent collaboration offers clear advantages in error recovery and reasoning diversity, but also introduces coordination overhead, latency, and reliance on critic agents. This is an active line of research because mitigating causal hallucination require not just a better model but a better reasoning process.

Summary. Now let us compare. RL reshapes model behavior through reward-driven post-training; neuro-symbolic methods make reasoning auditable; CoT externalizes intermediate logic for verification; and multi-

agent systems enable procedual error correction. Clearly, a promising direction is to combine verifier-trained rewards, structured representations, evidence-grounded reasoning traces, and lightweight agentic verification.

5.6 Cross-Modal Hallucination Mitigation

Cross-modal hallucination mitigation addresses failures in integrating evidence across visual, audio, and textual modalities. Their mitigation approaches fall into two subcategories: (i) *temporal synchronization and interleaved encoding*; and (ii) *cross-modal alignment and contrastive tuning*.

Temporal Synchronization and Interleaved Encoding Fine-grained temporal alignment between audio and visual signals is essential for preventing misattribution of auditory cues. Early Vid-LLMs such as Video-LLaMA (Zhang et al., 2023b) processed modalities through largely independent adapters, resulting in weak audio exploitation. VideoLLaMA 2 (Cheng et al., 2024a) replaced this with a Spatial-Temporal Convolution connector employing 3D convolutions and integrated a BEATs audio encoder, improving on EgoSchema but remaining coarse-grained beyond eight frames.

As an explicit temporal encoding method, OMCAT (Goel et al., 2024a) introduces Rotary Time Embeddings (RoTE) encoding absolute timestamps as rotation angles into both audio and visual tokens, as many cross-modal hallucinations arise from the model's inability to determine which signal belongs to which moment. OMCAT nearly doubled VideoLLaMA 2's accuracy and outperformed AVicuna (Shu et al., 2023).

Along this direction, Dolphin (Guo et al., 2025e) addresses both spatial and temporal dimensions through a multi-scale audio-visual adapter and interleaved temporal merging, reducing hallucination rates by nearly 40% on speech-aware samples. Temporal synchronization is most effective when paired with source localization. Chen et al. (2025b) shifted to frame-level alignment through a streaming paradigm densely interleaving ASR word tokens with video frames based on precise timestamps, achieving state-of-the-art real-time commentary with 30× latency reduction over Vid2Seq (Yang et al., 2023), though inheriting ASR errors. Ge et al. (2025) augmented captioning models with an auxiliary audio encoder and adaptive duration-based synchronization for multi-granularity temporally grounded captioning.

The central insight of these methods is: cross-modal hallucination often occurs asa sequencing problem rather as a representation problem. Models fail because they fuse modalities too early, too coarsely, or without a reliable notion of event order. A persistent challenge worth noting is how to handle temporally overlapping or weakly correlated audio-visual events.

Cross-Modal Alignment While temporal synchronization addresses *when* modalities should be aligned, cross-modal alignment addresses *what* should be aligned by enforcing representational consistency across video, audio, and text feature spaces.

Yang et al. (2024) proposed a self-supervised contrastive regularization framework that contrasts original and perturbed multimodal inputs with temporal order regularization, reducing reliance on superficial question-scene correlations beyond the invariant grounding of Li et al. (2022c). Bansal et al. (2023) leveraged contrast captions with temporal and action perturbations, improving temporal reasoning (Bagad et al., 2023; Momeni et al., 2023). Zhao et al. (2025b) replaced VideoCon's binary contrastive signal with a self-training hallucination correction objective, outperforming it by up to 17.9% on VELOCITI (Saravanan et al., 2025).

InternVideo2 (Wang et al., 2024k) introduced a progressive three-stage paradigm unifying masked video modeling, cross-modal contrastive learning across video-audio-speech-text, and next-token prediction within a 6B-parameter model. It goes beyond video-text contrastive scope (Li et al., 2023b) and achieves retrieval gains on MSR-VTT (Xu et al., 2016). Li et al. (2023c) introduced object-aware adaptive-positivity contrastive learning aligning question-object and audio-object pairs individually on MUSIC-AVQA (Li et al., 2022a). Clearly, alignment has become more about constructing a shared multimodal space preserving motion, speech, and semantics.

In the causal reasoning line of work, Chen et al. (2025d) formulated a Structural Causal Model for multimodal deconfounding, employing front-door adjustment on visual features and back-door adjustment on linguistic features. Causal alignment is better suited to reducing unfaithful grounding than correlation-based methods

alone, since a model may attend to aligned segments that are not the true basis for the answer. This reduced bias-induced errors over data-driven baselines (Xiao et al., 2023a), demonstrating that causal interventions complement contrastive strategies in weakly supervised settings.

There are also architectural innovations for alignment. Liu et al. (2023a) proposed a cascaded Q-Former reducing severe hallucination rates from 34% to 20% on Video-CSR (Liu et al., 2023b). Additional efforts include mutual knowledge transfer (Weng & Li, 2022), bidirectional set-level alignment (Wang et al., 2023b), multi-grained correspondence learning (Wang et al., 2024l), codebook-based discretization (Zhu et al., 2025), sparse masked pre-training (Lin et al., 2022), congruity matching (Zeng et al., 2023), and dual-task representations (Wang et al., 2024f). The most effective recent methods distinguish themselves by finer control over what is aligned: timestamps, objects, speech spans, event order, or causal evidence.

Summary. In the cross-modality category, temporal synchronization ensures correct chronological pairing, and cross-modal alignment enforces semantic consistency across feature spaces. Synchronization has evolved from independent fusion through timestamp encoding to fine-grained interleaving. Alignment spans contrastive regularization, multimodal pretraining, object-level grounding, and causal deconfounding. Fine-grained methods improve faithfulness but increase cost. Open challenges include robust reasoning over temporally overlapping or weakly correlated events, handling arbitrary modality combinations, and efficient scaling to long-form video.

6 Future Directions

We have mentioned many open challenges along the way, which limit the reliability of Vid-LLMs in real-world deployment. Here we highlight future directions we believe promising.

6.1 Benchmarking and Evaluation for Video Hallucination

Current benchmarks overwhelmingly measure task-level accuracy without disentangling whether strong performance reflects faithful video grounding or fluent confabulation (Maaz et al., 2023; Zhang et al., 2024c). Actually, a model may achieve highly competitive scores while still fabricating objects, inventing temporal orderings, or hallucinating causal relationships not evident in the video.

Designing benchmarks that directly evaluate hallucination in video understanding is challenging. First, annotations must specify not only what is present but also what is absent, i.e., explicit negative labels for objects, events, or temporal relations that models might plausibly hallucinate. Second, metrics must move beyond binary correctness to measure grounding accuracy, for example, whether each claim is supported by specific visual evidence at the correct temporal location (Xiao et al., 2023a; Guo et al., 2024b). Analyses on NExT-GQA (Xiao et al., 2024) show that models that achieve nearly 70% QA accuracy ground their answers in relevant segments only less than 20% of the time (Xiao et al., 2023a). Third, without saying, benchmarks must cover diverse temporal scales and hallucination types. Last but not least, rigorous benchmarks should move evaluation beyond only proxy tasks that may not reflect genuine faithfulness improvements.

6.2 Out-of-Schema Strategies

Most mitigation methods assume that encountered entities, actions, and event structures fall within the training distribution (Maaz et al., 2023; Chen et al., 2024c). But real-world videos routinely present novel entities and previously unseen event configurations. When confronted with such new inputs, Vid-LLMs may project unfamiliar percepts onto the nearest learned prototype and generate fluent but fabricated descriptions.

Vid-LLMs trained through next-token prediction or contrastive alignment maximize likelihood over a fixed semantic space, encouraging confident predictions even when evidence is ambiguous (Li et al., 2024c; Lin et al., 2023). Promising avenues along this line include explicit uncertainty estimation over visual concepts, shifting from categorical labeling toward grounded open-vocabulary descriptions composing outputs from primitive visual attributes (Wasim et al., 2023; Munasinghe et al., 2024), and large-scale open-world data expansion incorporating diverse real-world streams and long-tail event combinations. Also, enabling Vid-

LLMs to recognize the boundaries of their own knowledge would substantially reduce hallucinations that current closed-world methods leave unaddressed.

6.3 Temporal Causality and Action Dynamics via World-Modeling

Video-specific hallucinations stem not only from weak temporal modeling but also from insufficient physical-world understanding (Chang et al., 2025b; Tang et al., 2025f). Current Vid-LLMs are believed to lack explicit representations of object permanence, motion continuity, and cause-effect transitions, relying instead on appearance cues and language priors (Zhan et al., 2025b; Liu et al., 2024e). Future Vid-LLMs or Video World Models may explicitly model how the physical world evolves over time, learning latent world representations that maintain structured state information and predict state transitions under actions. Grounding reasoning in explicit world models could substantially reduce hallucinations related to action dynamics and false causal narratives.

7 Conclusion

Vid-LLMs and their capacities have rapidly advanced, but their reliability remains limited by hallucination, i.e., generating outputs not faithfully grounded in video evidence. Like and unlike hallucination in text or static images, video hallucination emerges from temporal reasoning, long-context information management, motion dynamics, and multimodal alignment.

This survey has presented a video-centric analysis of hallucination in Vid-LLMs. We introduce a unified taxonomy organizing video hallucination into six categories: temporal understanding, motion and action dynamics, long-context reasoning, object perception, causal reasoning, and cross-modal grounding. Then we systematically review mitigation techniques organized by the hallucinations they address.

Despite rapid progress, reliable video understanding requires further advances in hallucination-aware evaluation, open-world reasoning capabilities, and stronger modeling of the physical world. We hope this survey provides some help for researchers and practitioners working in the area of video understanding.

References

Daechul Ahn, Yura Choi, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. Tuning large multimodal models for videos using reinforcement learning from ai feedback. pp. 923–940, 2024.

Daechul Ahn, Yura Choi, Hyeonbeom Choi, Seongwon Cho, San Kim, and Jonghyun Choi. What happens when: Learning temporal orders of events in videos. *arXiv preprint arXiv:2512.08979*, 2025.

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Mingchen Zhuge, Jian Ding, Deyao Zhu, Jürgen Schmidhuber, and Mohamed Elhoseiny. Goldfish: Vision-language understanding of arbitrarily long videos. pp. 251–267, 2024.

Kyungho Bae, Jinhyung Kim, Sihaeng Lee, Soonyoung Lee, Gunhee Lee, and Jinwoo Choi. Mash-vlm: Mitigating action-scene hallucination in video-llms through disentangled spatial-temporal representations. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13744–13753, 2025.

Piyush Bagad, Makarand Tapaswi, and Cees G. M. Snoek. Test of time: Instilling video-language models with a sense of time. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2503–2516, 2023.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

Hritik Bansal, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, and Aditya Grover. Videocon: Robust video-language alignment via contrast captions. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13927–13937, 2023.

Xiaoyi Bao, Chen-Wei Xie, Hao Tang, Tingyu Weng, Xiaofeng Wang, Yun Zheng, and Xingang Wang. Dynimg: Key frames with visual prompts are good representation for multi-modal video understanding. *ArXiv*, abs/2507.15569, 2025.

Dionysia Danai Brilli, Dimitrios Mallis, Vassilis Pitsikalis, and P. Maragos. Ghr-vqa: Graph-guided hierarchical relational reasoning for video question answering. *ArXiv*, abs/2511.20201, 2025.

S. Buch, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. Flexible frame selection for efficient video reasoning. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29071–29082, 2025.

Kai-Po Chang, Wei-Yuan Cheng, Chi-Pin Huang, Fu-En Yang, and Yu-Chiang Frank Wang. Mitigating object and action hallucinations in multimodal llms via self-augmented contrastive alignment. *ArXiv*, abs/2512.04356, 2025a.

Kai-Po Chang, Wei-Yuan Cheng, Chi-Pin Huang, Fu-En Yang, and Yu-Chiang Frank Wang. Mitigating object and action hallucinations in multimodal llms via self-augmented contrastive alignment. *arXiv preprint arXiv:2512.04356*, 2025b.

Mustafa Chasmai, Gauri Jagatap, KV Gouthaman, Grant Van Horn, Subhransu Maji, and Andrea Fanelli. Moment sampling in video llms for long-form video qa. *ArXiv*, abs/2507.00033, 2025.

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. *ArXiv*, abs/2503.10200, 2025a.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18407–18418, 2024a.

Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. Live: Learning video llm with streaming speech transcription at scale. pp. 29083–29095, 2025b.

Junwen Chen, Jie Zhu, and Yu Kong. Atm: Action temporality modeling for video question answering. *Proceedings of the 31st ACM International Conference on Multimedia*, 2023a.

Kehua Chen. Chronoforge-rl: Chronological forging through reinforcement learning for enhanced video understanding. *ArXiv*, abs/2509.15800, 2025.

Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.

Lin Chen, Xilin Wei, Jinsong Li, Xiao wen Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. *ArXiv*, abs/2406.04325, 2024c.

Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos. *ArXiv*, abs/2408.14469, 2024d.

Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2159–2167, 2025c.

Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*, 2024e.

Sihan Chen, Xingjian He, Handong Li, Xiaojie Jin, Jiashi Feng, and J. Liu. Cosa: Concatenated sample pretrained vision-language foundation model. *ArXiv*, abs/2306.09085, 2023b.

Tieyuan Chen, Huabin Liu, Tianyao He, Yihang Chen, Chaofan Gan, Xiao Ma, Cheng Zhong, Yang Zhang, Yingxue Wang, Hui Lin, et al. Mecd: Unlocking multi-event causal discovery in video reasoning. *Advances in neural information processing systems*, 37:92554–92580, 2024f.

Weixing Chen, Yang Liu, Binglin Chen, Jiandong Su, Yongsen Zheng, and Liang Lin. Cross-modal causal relation alignment for video question grounding. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24087–24096, 2025d.

Xinlong Chen, Yuanxing Zhang, Yushuo Guan, Bohan Zeng, Yang Shi, Sihan Yang, Pengfei Wan, Qiang Liu, Liang Wang, and Tieniu Tan. Vidbridge-r1: Bridging qa and captioning for rl-based video understanding models with intermediate proxy tasks. 2025e.

Xueyi Chen, Keda Tao, Kele Shao, and Huan Wang. Streamingtom: Streaming token compression for efficient video understanding. *ArXiv*, abs/2510.18269, 2025f.

Yilong Chen, Xiang Bai, Zhibin Wang, Chengyu Bai, Yuhan Dai, Ming Lu, and Shanghang Zhang. Streamkv: Streaming video question-answering with segment-based kv cache retrieval and compression. *ArXiv*, abs/2511.07278, 2025g.

Yuefei Chen, Jiang Liu, Xiaodong Lin, and Ruixiang Tang. Countervqa: Evaluating and improving counterfactual reasoning in vision-language models for video understanding. *arXiv preprint arXiv:2511.19923*, 2025h.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. Scaling rl to long videos. *ArXiv*, abs/2507.07966, 2025i.

Yuyan Chen, Jiyuan Jia, Jiaxin Lu, Siyue Li, Yu Guan, Ming Yang, and Qingpei Guo. Vqaguider: Guiding multimodal large language models to answer complex video questions. pp. 7821–7834, 2025j.

Zhanwen Chen, Tianchun Wang, Yizhou Wang, Michal Kosinski, Xiang Zhang, Yun Fu, and Sheng Li. Through the theory of mind's eye: Reading minds with multimodal video large language models. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2025k.

Zhaoyu Chen, Hongnan Lin, Yongwei Nie, Fei Ma, Xuemiao Xu, Fei Yu, and Chengjiang Long. Invert4tvg: A temporal video grounding framework with inversion tasks for enhanced action understanding. *ArXiv*, abs/2508.07388, 2025l.

Zhawnen Chen, Tianchun Wang, Yizhou Wang, Michal Kosinski, Xiang Zhang, Yun Fu, and Sheng Li. Through the theory of mind's eye: Reading minds with multimodal video large language models. *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2024g.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Li Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *ArXiv*, abs/2406.07476, 2024a.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024b.

Md Intisar Chowdhury, Kittinun Aukkapinyo, Hiroshi Fujimura, Joo Ann Woo, Wasu Wasusatein, and Fadoua Ghourabi. Grid-logat: Grid based local and global area transcription for video question answering. *ArXiv*, abs/2505.24371, 2025.

Meng Chu, Yicong Li, and Tat seng Chua. *GraphVideoAgent: Enhancing Long-form Video Understanding with Entity Relation Graphs*. 2025.

Jihoon Chung, Tyler Zhu, Max Gonzalez Saez-Diez, Juan Carlos Niebles, Honglu Zhou, and Olga Russakovsky. Unifying specialized visual encoders for video language models. *ArXiv*, abs/2501.01426, 2025.

Ashim Dahal, Ankit Ghimire, Saydul Akbar Murad, and Nick Rahimi. Povqa: Preference-optimized video question answering with rationales for data efficiency. *ArXiv*, abs/2510.01009, 2025.

Wei Dai, Alan Luo, Zane Durante, Debadutta Dash, Arnold Milstein, Kevin Schulman, Ehsan Adeli, and Fei-Fei Li. Towards fine-grained video question answering. *ArXiv*, abs/2503.06820, 2025.

Jisheng Dang, Jingze Wu, Teng Wang, Xuanhui Lin, Nannan Zhu, Hongbo Chen, Wei-Shi Zheng, Meng Wang, and Tat-Seng Chua. Reinforcing video reasoning with focused thinking. *ArXiv*, abs/2505.24718, 2025a.

Jisheng Dang, Yizhou Zhang, Hao Ye, Teng Wang, Siming Chen, Huicheng Zheng, Yulan Guo, Jian-Huang Lai, and Bin Hu. Synpo: Synergizing descriptiveness and preference optimization for video detailed captioning. *ArXiv*, abs/2506.00835, 2025b.

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Zou, Kai-Wei Chang, and Wei Wang. Enhancing large vision language models with self-training on image comprehension. *Advances in Neural Information Processing Systems*, 37:131369–131397, 2024.

Xinpeng Ding, Kui Zhang, Jinahua Han, Lanqing Hong, Hang Xu, and Xiaomeng Li. Pami-vdpo: Mitigating video hallucinations by prompt-aware multi-instance video preference learning. *ArXiv*, abs/2504.05810, 2025.

Xinxin Dong, Baoyun Peng, Haokai Ma, Yufei Wang, Zixuan Dong, Fei Hu, and Xiaodong Wang. Leadqa: Llm-driven context-aware temporal grounding for video question answering. *ArXiv*, abs/2507.14784, 2025a.

Zixuan Dong, Baoyun Peng, Yufei Wang, Lin Liu, Xinxin Dong, Yunlong Cao, and Xiaodong Wang. See what you need: Query-aware visual intelligence through reasoning-perception loops. *ArXiv*, abs/2508.17932, 2025b.

Hazel Doughty, Fida Mohammad Thoker, and Cees G. M. Snoek. Locomotion: Learning motion-focused video-language representations. *ArXiv*, abs/2410.12018, 2024.

Shanshan Du and Hanli Wang. Scene-aware graph-enhanced multimodal collaboration for video-grounded dialogue. *2025 International Symposium on Machine Learning and Media Computing (MLMC)*, pp. 1–6, 2025.

Yipeng Du, Tiehan Fan, Kepan Nan, Rui Xie, Penghao Zhou, Xiang Li, Jian Yang, Zhenheng Yang, and Ying Tai. Motionsight: Boosting fine-grained motion understanding in multimodal llms. *ArXiv*, abs/2506.01674, 2025a.

Yipeng Du, Tiehan Fan, Kepan Nan, Rui Xie, Penghao Zhou, Xiang Li, Jian Yang, Zhenheng Yang, and Ying Tai. Motionsight: Boosting fine-grained motion understanding in multimodal llms. *arXiv preprint arXiv:2506.01674*, 2025b.

Chrisantus Eze and Christopher Crick. Learning by watching: A review of video-based learning approaches for robot manipulation. *IEEE Access*, 2025.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. *ArXiv*, abs/2403.11481, 2024.

Bo Fang, Yuxin Song, Qiangqiang Wu, Haoyuan Sun, Wenhao Wu, and Antoni B. Chan. Viss-r1: Self-supervised reinforcement video reasoning. *ArXiv*, abs/2511.13054, 2025.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong Li Lee, and W. Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. *ArXiv*, abs/2501.03230, 2024a.

Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:7701–7719, 2024b.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *ArXiv*, abs/2503.21776, 2025.

Yicheng Feng, Yijiang Li, Wanpeng Zhang, Sipeng Zheng, and Zongqing Lu. Videoorion: Tokenizing object dynamics in videos. *ArXiv*, abs/2411.16156, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Hong Gao, Yiming Bao, Xuezhan Tu, Bin Zhong, and Min-Ling Zhang. Apvr: Hour-level long video understanding with adaptive pivot visual information retrieval. *ArXiv*, abs/2506.04953, 2025a.

Hongcheng Gao, Jiashu Qu, Jingyi Tang, Baolong Bi, Yue Liu, Hongyu Chen, Li Liang, Li Su, and Qingming Huang. Exploring hallucination of large multimodal models in video understanding: Benchmark, analysis and mitigation. *arXiv preprint arXiv:2503.19622*, 2025b.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pp. 5267–5275, 2017.

Shida Gao, Feng Xue, Xiangfeng Wang, Anlong Ming, Teng Long, Yihua Shao, Haozhe Wang, Zhaowen Lin, Wei Wang, and Niculae Sebe. 1 + 1 > 2: Detector-empowered video large language model for spatio-temporal grounding and reasoning. *ArXiv*, abs/2512.06673, 2025c.

Yuying Ge, Yizhuo Li, Yixiao Ge, and Ying Shan. Divot: Diffusion powers video tokenizer for comprehension and generation. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13606–13617, 2024.

Yuying Ge, Yixiao Ge, Chen Li, Teng Wang, Junfu Pu, Yizhuo Li, Lu Qiu, Jin Ma, Lisheng Duan, Xinyu Zuo, Jinwen Luo, Weibo Gu, Zexuan Li, Xiaojing Zhang, Yang-Dan Tao, Han Hu, Di Wang, and Ying Shan. Arc-hunyuan-video-7b: Structured video comprehension of real-world shorts. *ArXiv*, abs/2507.20939, 2025.

Arushi Goel, Karan Sapra, Matthieu Le, Rafael Valle, Andrew Tao, and Bryan Catanzaro. Omcat: Omni context aware transformer. *ArXiv*, abs/2410.12109, 2024a.

Arushi Goel, Karan Sapra, Matthieu Le, Rafael Valle, Andrew Tao, and Bryan Catanzaro. Omcat: Omni context aware transformer. *arXiv preprint arXiv:2410.12109*, 2024b.

Andreas Goulas, V. Mezaris, and Ioannis Patras. Vidctx: Context-aware video question answering with image models. *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2024.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11287–11297, 2021.

Xin Gu, Yaojie Shen, Chenxi Luo, Tiejian Luo, Yan Huang, Yuewei Lin, Heng Fan, and Libo Zhang. Knowing your target: Target-aware transformer makes better spatio-temporal video grounding. *ArXiv*, abs/2502.11168, 2025.

Bhanu Prakash Reddy Guda, Tanmay Kulkarni, Adithya Sampath, and Swarnashree Mysore Sathyendra. Causal understanding for video question answering. *arXiv preprint arXiv:2407.20257*, 2024.

Weiyu Guo, Ziyang Chen, Shaoguang Wang, Jianxiang He, Yijie Xu, Jinhui Ye, Ying Sun, and Hui Xiong. Logic-in-frames: Dynamic keyframe search via visual semantic-logical verification for long video understanding. *ArXiv*, abs/2503.13139, 2025a.

Yanan Guo, Wenhui Dong, Jun Song, Shiding Zhu, Xuan Zhang, Hanqing Yang, Yingbo Wang, Yang Du, Xianing Chen, and Bo Zheng. Fila-video: Spatio-temporal compression for fine-grained long video understanding. *ArXiv*, abs/2504.20384, 2025b.

Yongxin Guo, Jingyu Liu, Mingda Li, Xiaoying Tang, Xi Chen, and Bo Zhao. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. *ArXiv*, abs/2405.13382, 2024a.

Yongxin Guo, Jingyu Liu, Mingda Li, Xiaoying Tang, Qingbin Liu, and Xi Chen. Trace: Temporal grounding video llm via causal event modeling. *ArXiv*, abs/2410.05643, 2024b.

Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. pp. 3302–3310, 2025c.

Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3302–3310, 2025d.

Yuxin Guo, Shuailei Ma, Shijie Ma, Xiaoyi Bao, Chen-Wei Xie, Kecheng Zheng, Tingyu Weng, Siyang Sun, Yun Zheng, and Wei Zou. Aligned better, listen better for audio-visual large language models. *arXiv*, abs/2504.02061, 2025e.

Yuxin Guo, Shuailei Ma, Shijie Ma, Xiaoyi Bao, Chen-Wei Xie, Kecheng Zheng, Tingyu Weng, Siyang Sun, Yun Zheng, and Wei Zou. Aligned better, listen better for audio-visual large language models. *arXiv preprint arXiv:2504.02061*, 2025f.

Antoine Hanna-Asaad, Decky Aspandi, and Titus Zaharia. Multi-modal interpretable automatic video captioning. *arXiv preprint arXiv:2411.06872*, 2024.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13504–13514, 2024.

Haibin He, Qihuang Zhong, Juhua Liu, Bo Du, Peng Wang, and Jing Zhang. Sfa: Scan, focus, and amplify toward guidance-aware answering for video textvqa. *ArXiv*, abs/2511.20190, 2025a.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. Framethinker: Learning to think with long videos via multi-turn frame spotlighting. *ArXiv*, abs/2509.24304, 2025b.

Miran Heo, Min-Hung Chen, De-An Huang, Sifei Liu, Subhashree Radhakrishnan, Seon Joo Kim, Yu-Chiang Frank Wang, and Ryo Hachiuma. Omni-rgpt: Unifying image and video region-level understanding via token marks. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3919–3930, 2025.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *CVPR*, pp. 8450–8460, 2025a.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8450–8460, 2025b.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, et al. M-llm based video frame selection for efficient video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13702–13712, 2025a.

Yangliu Hu, Zikai Song, Na Feng, Yawei Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Sf2t: Self-supervised fragment finetuning of video-llms for fine-grained understanding. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29108–29117, 2025b.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14271–14280, 2023.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14271–14280, 2024a.

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. *ArXiv*, abs/2403.19046, 2024b.

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *European Conference on Computer Vision*, pp. 202–218. Springer, 2024c.

Haojian Huang, Hao Chen, Shengqiong Wu, Meng Luo, Jinlan Fu, Xinya Du, Hanwang Zhang, and Hao Fei. Vistadpo: Video hierarchical spatial-temporal direct preference optimization for large video models. *ArXiv*, abs/2504.13122, 2025a.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55, 2025b.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*, 2024d.

Zeyi Huang, Yuyang Ji, Xiaofang Wang, Nikhil Mehta, Tong Xiao, Donghyun Lee, Sigmund Vanvalkenburgh, Shengxin Zha, Bolin Lai, Licheng Yu, Ning Zhang, Yong Jae Lee, and Miao Liu. Building a mind palace: Structuring environment-grounded semantic graphs for effective long video analysis with llms. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24169–24179, 2025c.

Zhe Huang, Hao Wen, Aiming Hao, Bingze Song, Meiqi Wu, Jiahong Wu, Xiangxiang Chu, Sheng Lu, and Haoqian Wang. Taming hallucinations: Boosting mllms' video understanding via counterfactual video generation. *arXiv preprint arXiv:2512.24271*, 2025d.

Yoichi Ishibashi and Yoshimasa Nishimura. Self-organized agents: A llm multi-agent framework toward ultra large-scale code generation and optimization. *arXiv preprint arXiv:2404.02183*, 2024.

Md Mohaiminul Islam, Tushar Nagarajan, Huiyu Wang, Gedas Bertasius, and Lorenzo Torresani. Bimba: Selective-scan compression for long-range video question answering. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29096–29107, 2025.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2758–2766, 2017.

Pu Jian, Junhong Wu, Wei Sun, Chen Wang, Shuo Ren, and Jiajun Zhang. Look again, think slowly: Enhancing visual reflection in vision-language models. *ArXiv*, abs/2509.12132, 2025.

Jindong Jiang, Xiuyu Li, Zhijian Liu, Muyang Li, Guo Chen, Zhiqi Li, De-An Huang, Guilin Liu, Zhiding Yu, Kurt Keutzer, Sungjin Ahn, Jan Kautz, Hongxu Yin, Yao Lu, Song Han, and Wonmin Byeon. Token-efficient long video understanding for multimodal llms. *ArXiv*, abs/2503.04130, 2025a.

Yifan Jiang, Yueying Wang, Rui Zhao, Toufiq Parag, Zhimin Chen, Zhenyu Liao, and Jayakrishnan Unnikrishnan. Videop2r: Video understanding from perception to reasoning. *arXiv preprint arXiv:2511.11113*, 2025b.

Yuanyuan Jiang and Jianqin Yin. Target-aware spatio-temporal reasoning via answering questions in dynamics audio-visual scenarios. *arXiv preprint arXiv:2305.12397*, 2023.

Hongbo Jin, Jiayu Ding, Siyi Xie, Guibo Luo, and Ge Li. Vista: Mitigating semantic inertia in video-llms via training-free dynamic chain-of-thought routing. 2025a.

Hongbo Jin, Ruyang Liu, Wenhao Zhang, Guibo Luo, and Ge Li. Cot-vid: Dynamic chain-of-thought routing with self verification for training-free video reasoning. *arXiv preprint arXiv:2505.11830*, 2025b.

Hongbo Jin, Qingyuan Wang, Wenhao Zhang, Yang Liu, and Sijie Cheng. Videomem: Enhancing ultra-long video understanding via adaptive memory management. *ArXiv*, abs/2512.04540, 2025c.

Hongbo Jin, Qingyuan Wang, Wenhao Zhang, Yang Liu, and Sijie Cheng. Videomem: Enhancing ultra-long video understanding via adaptive memory management. *arXiv preprint arXiv:2512.04540*, 2025d.

Yang Jin, Zhicheng Sun, Kun Xu, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, Kun Gai, and Yadong Mu. Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. *ArXiv*, abs/2402.03161, 2024.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024.

Minjoon Jung, Junbin Xiao, Byoung-Tak Zhang, and Angela Yao. On the consistency of video large language models in temporal comprehension. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13713–13722, 2024.

Minjoon Jung, Junbin Xiao, Byoung-Tak Zhang, and Angela Yao. On the consistency of video large language models in temporal comprehension. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13713–13722, 2025.

Y. Jung, Minsoo Kim, Seungtaek Choi, Jihyuk Kim, Minji Seo, and Seung won Hwang. Retrieval-augmented video encoding for instructional captioning. pp. 8554–8568, 2023.

Umihiro Kamoto, Tatsuya Ishibashi, and Noriyuki Kugo. Dive: Deep-search iterative video exploration a technical report for the cvrr challenge at cvpr 2025. *ArXiv*, abs/2506.21891, 2025.

Evangelos Kazakos, Cordelia Schmid, and Josef Sivic. Large-scale pre-training for grounded video caption generation. *ArXiv*, abs/2503.10781, 2025a.

Evangelos Kazakos, Cordelia Schmid, and Josef Sivic. Large-scale pre-training for grounded video caption generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24434–24444, 2025b.

Aisha Urooj Khan, Hilde Kuehne, Bo Wu, Kim Chheu, Walid Bousselham, Chuang Gan, N. Lobo, and M. Shah. Learning situation hyper-graphs for video question answering. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14879–14889, 2023.

Younggun Kim, Ahmed S Abdelrahman, and Mohamed Abdel-Aty. Vru-accident: A vision-language benchmark for video question answering and dense captioning for accident scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 761–771, 2025.

Fanqi Kong, Weiqin Zu, Xinyu Chen, Yaodong Yang, Song-Chun Zhu, and Xue Feng. Siv-bench: A video benchmark for social interaction understanding and reasoning. *arXiv preprint arXiv:2506.05425*, 2025.

Yogesh Kulkarni and Pooyan Fazli. Videosavi: Self-aligned video language models without human supervision. *ArXiv*, abs/2412.00624, 2024.

Yogesh Kulkarni and Pooyan Fazli. Videopasta: 7k preference pairs that matter for video-llm alignment. *ArXiv*, abs/2504.14096, 2025.

Ji Soo Lee, Jongha Kim, Jeehye Na, Jinyoung Park, and Hyunwoo J. Kim. Vidchain: Chain-of-tasks with metric-based direct preference optimization for dense video captioning. pp. 4499–4507, 2025.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13723–13733, 2025a.

Chaoyu Li, Tianzhi Li, Fei Tao, Zhenyu Zhao, Ziqian Wu, Maozheng Zhao, Juntong Song, Cheng Niu, and Pooyan Fazli. Frameoracle: Learning what to see and how much to see in videos. *ArXiv*, abs/2510.03584, 2025b.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19108–19118, 2022a.

Hongyu Li, Songhao Han, Yue Liao, Junfeng Luo, Jialin Gao, Shuicheng Yan, and Si Liu. Reinforcement learning tuning for videollms: Reward design and data efficiency. *ArXiv*, abs/2506.01908, 2025c.

Jiahua Li, Kun-Juan Wei, Zhe Xu, Zibo Su, Xu Yang, and Cheng Deng. Perceive, reflect and understand long video: Progressive multi-granular clue exploration with interactive agents. *ArXiv*, abs/2509.24943, 2025d.

Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21273–21282, 2022b.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11963–11974, 2023a.

Jiaze Li, Yaya Shi, Zongyang Ma, Haoran Xu, Feng Cheng, Huihui Xiao, Ruiwen Kang, Fan Yang, Tingting Gao, and Di Zhang. imove: Instance-motion-aware video understanding. pp. 23959–23975, 2025e.

Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19891–19903, 2023b.

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

Lei Li, Yuanxin Liu, Linli Yao, Peiyuan Zhang, Chenxin An, Lean Wang, Xu Sun, Lingpeng Kong, and Qi Liu. Temporal reasoning transfer from text to video. *ArXiv*, abs/2410.06166, 2024b.

Rui Li, Xiaohan Wang, Yuhui Zhang, Zeyu Wang, and S. Yeung-Levy. Temporal preference optimization for long-form video understanding. *ArXiv*, abs/2501.13919, 2025f.

Wei Li, Hehe Fan, Yongkang Wong, Mohan S. Kankanhalli, and Yi Yang. Topa: Extending large language models for video understanding via text-only pre-alignment. *Advances in Neural Information Processing Systems 37*, 2024c.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024d.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyun Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *ArXiv*, abs/2504.06958, 2025g.

Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat seng Chua. Invariant grounding for video question answering. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2918–2927, 2022c.

Yicong Li, Xiang Wang, Junbin Xiao, and Tat seng Chua. *Equivariant and Invariant Grounding for Video Question Answering*. 2022d.

Yixuan Li, Changli Tang, Jimin Zhuang, Yudong Yang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. Improving llm video understanding with 16 frames per second. *ArXiv*, abs/2503.13956, 2025h.

Yunheng Li, Jing Cheng, Shaoyong Jia, Hangyi Kuang, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. Tempsamp-r1: Effective temporal sampling with reinforcement fine-tuning for video llms. *ArXiv*, abs/2509.18056, 2025i.

Yunxin Li, Xinyu Chen, Zitao Li, Zhenyu Liu, Longyue Wang, Wenhan Luo, Baotian Hu, and Min Zhang. Veripo: Cultivating long reasoning in video-llms via verifier-gudied iterative policy optimization. *ArXiv*, abs/2505.19000, 2025j.

Zhangbin Li, Dan Guo, Jinxing Zhou, Jing Zhang, and Meng Wang. Object-aware adaptive-positivity learning for audio-visual question answering. pp. 3306–3314, 2023c.

Zhangbin Li, Dan Guo, Jinxing Zhou, Jing Zhang, and Meng Wang. Object-aware adaptive-positivity learning for audio-visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3306–3314, 2024e.

Zinuo Li, Xian Zhang, Yongxin Guo, Mohammed Bennamoun, Farid Boussaid, Girish Dwivedi, Luqi Gong, and Qiuhong Ke. Watch and listen: Understanding audio-visual-speech moments with multimodal LLM. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025k. URL `https://openreview.net/forum?id=kWGJa9ZO3M`.

Jianxin Liang, Xiaojun Meng, Huishuai Zhang, Yueqian Wang, Jiansheng Wei, and Dongyan Zhao. Reasvqa: Advancing videoqa with imperfect reasoning process. pp. 1696–1709, 2025.

Lili Liang and Guanglu Sun. Leveraging static relationships for intra-type and inter-type message passing in video question answering. *ArXiv*, abs/2504.02417, 2025.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *ArXiv*, abs/2311.10122, 2023.

Yuanze Lin, Chen Wei, Huiyu Wang, A. Yuille, and Cihang Xie. Smaug: Sparse masked autoencoder for efficient video-language pre-training. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2459–2469, 2022.

Sergey Linok, Vadim Semenov, Anastasia Trunova, Oleg Bulichev, and Dmitry Yudin. Dygenc: Encoding a sequence of textual scene graphs to reason and answer questions in dynamic scenes. *ArXiv*, abs/2505.03581, 2025.

Caihua Liu, Xu Li, Wenjing Xue, Wei Tang, and Xia Feng. Capturing rich behavior representations: A dynamic action semantic-aware graph transformer for video captioning. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025a.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024a.

Haogeng Liu, Qihang Fan, Tingkai Liu, Linjie Yang, Yunzhe Tao, Huaibo Huang, Ran He, and Hongxia Yang. Video-teller: Enhancing cross-modal generation with fusion and decoupling. *arXiv*, abs/2310.04991, 2023a.

Huabin Liu, Filip Ilievski, and Cees G. M. Snoek. Commonsense video question answering through video-grounded entailment tree reasoning. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3262–3271, 2025b.

Hui Liu and Xiaojun Wan. *QAVidCap: Enhancing Video Captioning through Question Answering Techniques*. 2024.

Jin Liu, Guoxiang Wang, Jialong Xie, F. Zhou, and Huijuan Xu. Video question answering with semantic disentanglement and reasoning. *IEEE Transactions on Circuits and Systems for Video Technology*, 34: 3663–3673, 2024b.

Keliang Liu, Dingkang Yang, Ziyun Qian, Weijie Yin, Yuchi Wang, Hongsheng Li, Jun Liu, Peng Zhai, Yang Liu, and Lihua Zhang. Reinforcement learning meets large language models: A survey of advancements and applications across the llm lifecycle. *arXiv preprint arXiv:2509.16679*, 2025c.

Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners. *ArXiv*, abs/2404.00308, 2024c.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. Bolt: Boost large vision-language model without training for long-form video understanding. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3318–3327, 2025d.

Tingkai Liu, Yunzhe Tao, Haogeng Liu, Qihang Fan, Ding Zhou, Huaibo Huang, Ran He, and Hongxia Yang. Video-csr: Complex video digest creation for visual-language models. 2023b.

Xiangrui Liu, Yan Shu, Zheng Liu, Ao Li, Yang Tian, and Bo Zhao. Video-xl-pro: Reconstructive token compression for extremely long video understanding. *ArXiv*, abs/2503.18478, 2025e.

Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:11624–11641, 2022a.

Yang Liu, Ying Tan, Jingzhou Luo, and Weixing Chen. Vcd: Visual causality discovery for cross-modal question reasoning. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 309–322. Springer, 2023c.

Yang Liu, Pengxiang Ding, Siteng Huang, Min Zhang, Han Zhao, and Donglin Wang. Pite: Pixel-temporal alignment for large video-language model. *ArXiv*, abs/2409.07239, 2024d.

Yang Liu, Pengxiang Ding, Siteng Huang, Min Zhang, Han Zhao, and Donglin Wang. Pite: Pixel-temporal alignment for large video-language model. In *European Conference on Computer Vision*, pp. 160–176. Springer, 2024e.

Yi Liu, Limin Wang, Yali Wang, Xiao Ma, and Yu Qiao. Fineaction: A fine-grained video dataset for temporal action localization. *IEEE transactions on image processing*, 31:6937–6950, 2022b.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024f.

Hao Lu, Jiahao Wang, Yaolun Zhang, Ruohui Wang, Xuanyu Zheng, Yepeng Tang, Dahua Lin, and Lewei Lu. Elv-halluc: Benchmarking semantic aggregation hallucinations in long video understanding. *arXiv preprint arXiv:2508.21496*, 2025.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, et al. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*, 2025a.

Romy Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. When thinking drifts: Evidential grounding for robust video reasoning. *ArXiv*, abs/2510.06077, 2025b.

Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Ming-Hui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *ArXiv*, abs/2306.07207, 2023.

Yongdong Luo, Xiawu Zheng, Xiao Yang, Guilin Li, Haojia Lin, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *ArXiv*, abs/2411.13093, 2024.

Fabio Lusha, Agnese Chiatti, Sara Pidò, Nico Catalano, and Matteo Matteucci. Graph against the machine: Neuro-symbolic approach for enhanced video question answering. In *Proceedings of the workshop on Advanced Neuro-symbolic Applications (ANSyA) at ECAI*, 2025.

Chenyang Lyu, Tianbo Ji, Yvette Graham, and Jennifer Foster. Semantic-aware dynamic retrospective-prospective reasoning for event-level video question answering. pp. 50–56, 2023.

Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. Vista-llama: Reducing hallucination in video language models via equal distance to visual tokens. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13151–13160, 2023.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18936–18946, 2024.

Muhammad Maaz, H. Rasheed, Salman H. Khan, and F. Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. pp. 12585–12602, 2023.

Muhammad Maaz, H. Rasheed, F. Khan, and Salman H. Khan. Video-r2: Reinforcing consistent and grounded reasoning in multimodal language models. 2025.

Neelu Madan, Andreas Møgelmose, Rajat Modi, Yogesh S Rawat, and Thomas B Moeslund. Foundation models for video understanding: A survey. *arXiv preprint arXiv:2405.03770*, 2024.

S. Malik, Moyuru Yamada, Ayush Singh, and Dishank Aggarwal. Ravu: Retrieval augmented video understanding with compositional reasoning over graph. *ArXiv*, abs/2505.03173, 2025.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36: 46212–46244, 2023.

Jianguo Mao, Wenbin Jiang, Xiangdong Wang, Zhifan Feng, Yajuan Lyu, Hong Liu, and Yong Zhu. Dynamic multistep reasoning based on video scene graph for video question answering. In *North American Chapter of the Association for Computational Linguistics*, pp. 3894–3904, 2022.

Mihai Masala and Marius Leordeanu. From vision to language through graph of events in space and time: An explainable self-supervised approach. *ArXiv*, abs/2507.04815, 2025.

Desen Meng, Rui Huang, Zhilin Dai, Xinhao Li, Yifan Xu, Jun Zhang, Zhenpeng Huang, Meng Zhang, Ling-Kun Zhang, Yi Liu, and Limin Wang. Videocap-r1: Enhancing mllms for video captioning via structured thinking. *ArXiv*, abs/2506.01725, 2025.

Sachit Menon, Ahmet Iscen, Arsha Nagrani, Tobias Weyand, Carl Vondrick, and Cordelia Schmid. Caviar: Critic-augmented video agentic reasoning. *ArXiv*, abs/2509.07680, 2025.

Juhong Min, S. Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13235–13245, 2024.

Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and C. Schmid. Verbs in action: Improving verb understanding in video-language models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15533–15545, 2023.

Tony Montes and Fernando Lozano. Viqagent: Zero-shot video question answering via agent with open-vocabulary grounding validation. *ArXiv*, abs/2505.15928, 2025.

Saman Motamed, Minghao Chen, L. V. Gool, and Iro Laina. Travl: A recipe for making video-language models better judges of physics implausibility. *ArXiv*, abs/2510.07550, 2025.

Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, H. Rasheed, Salman H. Khan, Mubarak Shah, and F. Khan. Pg-video-llava: Pixel grounding large video-language models. *ArXiv*, abs/2311.13435, 2023.

Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric P. Xing, F. Khan, and Salman H. Khan. Videoglamm : A large multimodal model for pixel-level visual grounding in videos. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19036–19046, 2024.

Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See Kiong Ng, and Luu Anh Tuan. Video-language understanding: A survey from model architecture, model training, and data perspectives. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3636–3657, 2024.

Hyunjong Ok and Jaeho Lee. Do video language models really know where to look? diagnosing attention failures in video language models. *ArXiv*, abs/2509.01167, 2025.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025a.

Kun Ouyang, Yuanxin Liu, Linli Yao, Yishuo Cai, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Conan: Progressive learning to reason like a detective over multi-scale visual evidence. *ArXiv*, abs/2510.20470, 2025b.

Junwen Pan, Qizhe Zhang, Rui Zhang, Ming Lu, Xin Wan, Yuan Zhang, Chang Liu, and Qi She. Timesearch-r: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning. *ArXiv*, abs/2511.05489, 2025.

Ziqi Pang and Yu-Xiong Wang. Mr. video: "mapreduce" is the principle for long video understanding. *ArXiv*, abs/2504.16082, 2025.

Paritosh Parmar, Eric Peh, and Basura Fernando. Chainreaction: Causal chain-guided reasoning for modular and explainable causal-why video question answering. 2025.

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

Min Peng, Chongyang Wang, Yuan Gao, Yu Shi, and Xiang-Dong Zhou. Multilevel hierarchical network with multiscale sampling for video question answering. pp. 1276–1282, 2022.

Wenshuo Peng, Gongxuan Wang, Tianmeng Yang, Chuanhao Li, Xiaojie Xu, Hui He, and Kaipeng Zhang. Svbench: Evaluation of video generation models on social reasoning. *arXiv preprint arXiv:2512.21507*, 2025.

Chiara Plizzari, A. Tonioni, Yongqin Xian, Achin Kulshrestha, and Federico Tombari. Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24129–24138, 2025.

Tobia Poppi, Burak Uzkent, Amanmeet Garg, Lucas Porto, Garin Kessler, Yezhou Yang, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara, and Florian Schiffers. Countervid: Counterfactual video generation for mitigating action and temporal hallucinations in video-language models. *arXiv preprint arXiv:2601.04778*, 2026.

Shraman Pramanick, Effrosyni Mavroudi, Yale Song, Rama Chellappa, Lorenzo Torresani, and Triantafyllos Afouras. Enrich and detect: Video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24297–24308, 2025.

Minghao Qin, Xiangrui Liu, Zhengyang Liang, Yan Shu, Huaying Yuan, Junjie Zhou, Shitao Xiao, Bo Zhao, and Zheng Liu. Video-xl-2: Towards very long-video understanding through task-aware kv sparsification. *ArXiv*, abs/2506.19225, 2025a.

Minghao Qin, Yan Shu, Peitian Zhang, Kun Lun, Huaying Yuan, Junjie Zhou, Shitao Xiao, Bo Zhao, and Zheng Liu. Task-aware kv compression for cost-effective long video understanding. *ArXiv*, abs/2506.21184, 2025b.

Haiyi Qiu, Minghe Gao, Long Qian, Kaihang Pan, Qifan Yu, Juncheng Li, Wenjie Wang, Siliang Tang, Yueting Zhuang, and Tat-Seng Chua. Step: Enhancing video-llms' compositional reasoning by spatio-temporal graph-guided self-training. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3284–3294, 2024.

Tianheng Qiu, Jingchun Gao, Jingyu Li, Hui Yi Leong, Xuan Huang, Xi Wang, Xiaocheng Zhang, Kele Xu, and Lan Zhang. *IntentVCNet: Bridging Spatio-Temporal Gaps for Intention-Oriented Controllable Video Captioning*. 2025.

Tingyu Qu, Mingxiao Li, T. Tuytelaars, and M. Moens. Ts-llava: Constructing visual tokens through thumbnail-and-sampling for training-free video large language models. *ArXiv*, abs/2411.11066, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and M. Ryoo. Understanding long videos with multimodal language models. 2024.

Ali Rasekh, Erfan Bagheri Soula, Omid Daliran, Simon Gottschalk, and Mohsen Fayyaz. Enhancing temporal understanding in video-llms through stacked temporal attention in vision encoders. *ArXiv*, abs/2510.26027, 2025.

H. Rasheed, Mohammed Zumri, Muhammad Maaz, Ming-Hsuan Yang, F. Khan, and Salman H. Khan. Video-com: Interactive video reasoning via chain of manipulations. *ArXiv*, abs/2511.23477, 2025.

Ruchit Rawal, Reza Shirkavand, Heng Huang, Gowthami Somepalli, and Tom Goldstein. Argus: Hallucination and omission evaluation in video-llms. *arXiv preprint arXiv:2506.07371*, 2025.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323, 2023.

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

Weiming Ren, Huan Yang, Jie Min, Cong Wei, and Wenhu Chen. Vista: Enhancing long-duration and high-resolution video understanding by video spatiotemporal augmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3804–3814, 2024b.

Weiming Ren, Wentao Ma, Huan Yang, Cong Wei, Ge Zhang, and Wenhu Chen. Vamba: Understanding hour-long videos with hybrid mamba-transformers. *ArXiv*, abs/2503.11579, 2025.

Hector Garcia Rodriguez, Boris Meinardus, Anil Batra, Anna Rohrbach, and Marcus Rohrbach. Chrono: A simple blueprint for representing time in mllms. 2024.

Sai Bhargav Rongali, Mohamad Hassan N C, Ankit Jha, Neha Bhargava, Saurabh Prasad, and Biplab Banerjee. Foundation models and adaptive feature selection: A synergistic approach to video question answering. *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 9269–9279, 2024.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. A comprehensive survey of hallucination in large language, image, video and audio foundation models. *arXiv preprint arXiv:2405.09589*, 2024.

Kate Sanders, Nathaniel Weir, and Benjamin Van Durme. Tv-trees: Multimodal entailment trees for neuro-symbolic video reasoning. *ArXiv*, abs/2402.19467, 2024.

Saul Santos, António Farinhas, Daniel C. McNamee, and André F. T. Martins. ∞-video: A training-free approach to long video understanding via continuous-time memory consolidation. *ArXiv*, abs/2501.19098, 2025.

Darshana Saravanan, Varun Gupta, Darshan Singh, Zeeshan Khan, Vineet Gandhi, and Makarand Tapaswi. Velociti: Benchmarking video-language compositional reasoning with strict entailment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18914–18924, 2025.

Ashish Seth, Utkarsh Tyagi, Ramaneswaran Selvakumar, Nishit Anand, Sonal Kumar, Sreyan Ghosh, Ramani Duraiswami, Chirag Agarwal, and Dinesh Manocha. Egoillusion: Benchmarking hallucinations in egocentric video understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28449–28468, 2025.

M. G. Shankar and D. Surendran. An effective video captioning based on language description using a novel graylag deep kookaburra reinforcement learning. *EURASIP Journal on Image and Video Processing*, 2025, 2025.

Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2616–2625, 2020.

Xiaoqian Shen, Wenxuan Zhang, Jun Chen, and Mohamed Elhoseiny. Vgent: Graph-based retrieval-reasoning-augmented generation for long video understanding. *ArXiv*, abs/2510.14032, 2025.

Xuyang Shen, Dong Li, Jinxing Zhou, Zhen Qin, Bowen He, Xiaodong Han, Aixuan Li, Yuchao Dai, Lingpeng Kong, Meng Wang, et al. Fine-grained audible video description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10585–10596, 2023.

Derek Shi, Ruben Glatt, Christine Klymko, Shubham Mohole, Hongjun Choi, Shashank Kushwaha, Sam Sakla, and Felipe Leno da Silva. Oracle-rlaif: An improved fine-tuning framework for multi-modal video models through reinforcement learning from ranking feedback. *ArXiv*, abs/2510.02561, 2025.

Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. Enhancing video-llm reasoning via agent-of-thoughts distillation. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8523–8533, 2024.

Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. Audio-visual llm for video understanding. *ArXiv*, abs/2312.06720, 2023.

Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26160–26169, 2024.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26160–26169, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tianbo Ye, Yang Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232, 2023.

Boyuan Sun, Jiaxin Zhao, Xihan Wei, and Qibin Hou. Llava-scissor: Token compression with semantic connected components for video llms. *ArXiv*, abs/2506.21862, 2025a.

Guangyu Sun, Archit Singhal, Burak Uzkent, Mubarak Shah, Chen Chen, and Garin Kessler. From frames to clips: Efficient key clip selection for long-form video understanding. *ArXiv*, abs/2510.02262, 2025b.

Guangzhi Sun, Yixuan Li, Xiaodong Wu, Yudong Yang, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn s: Streaming audio-visual llms beyond length limits via memory. *ArXiv*, abs/2510.11129, 2025c.

Yiwei Sun, Zhihang Liu, Chuanbin Liu, Bowei Pu, Zhihan Zhang, and Hongtao Xie. Hallucination mitigation prompts long-term video understanding. *ArXiv*, abs/2406.11333, 2024.

Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. Long-form video-language pre-training with multimodal temporal contrastive learning. *ArXiv*, abs/2210.06031, 2022.

Agnese Taluzzi, Davide Gesualdi, Riccardo Santambrogio, Chiara Plizzari, Francesca Palermo, S. Mentasti, and Matteo Matteucci. From pixels to graphs: using scene and knowledge graphs for hd-epic vqa challenge. *ArXiv*, abs/2506.08553, 2025.

Wenhui Tan, Ruihua Song, Jiaze Li, Jianzhong Ju, and Zhenbo Luo. Think-clip-sample: Slow-fast frame selection for video understanding. *arXiv preprint arXiv:2601.11359*, 2026.

Canhui Tang, Zifan Han, Hongbo Sun, Sanpin Zhou, Xuchong Zhang, Xin Wei, Ye Yuan, Huayu Zhang, Jinglin Xu, and Hao Sun. Tspo: Temporal sampling policy optimization for long-form video language understanding. *ArXiv*, abs/2508.04369, 2025a.

Jiahao Tang, Jianguo Hu, Wenjun Huang, Shengzhi Shen, Jiakai Pan, Deming Wang, and Yanyu Ding. Spatio-temporal graph convolution transformer for video question answering. *IEEE Access*, 12:131664–131680, 2024.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128, 2025b.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025c.

Yunlong Tang, Daiki Shimada, Jing Bi, Mingqian Feng, Hang Hua, and Chenliang Xu. Empowering llms with pseudo-untrimmed videos for audio-visual temporal understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7293–7301, 2025d.

Zitian Tang, Shijie Wang, Junho Cho, Jaewook Yoo, and Chen Sun. How can objects help video-language understanding? *ArXiv*, abs/2504.07454, 2025e.

Zitian Tang, Shijie Wang, Junho Cho, Jaewook Yoo, and Chen Sun. How can objects help video-language understanding? *arXiv preprint arXiv:2504.07454*, 2025f.

Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249, 2021.

Sicheng Tao, Jungang Li, Yibo Yan, Junyan Zhang, Yubo Gao, Hanqian Li, Shuhang Xun, Yuxuan Fan, Hong Chen, Jianxiang He, and Xuming Hu. Moss-chatv: Reinforcement learning with process reasoning reward for video temporal reasoning. *ArXiv*, abs/2509.21113, 2025.

George Tom, Minesh Mathew, Sergi Garcia-Bordils, Dimosthenis Karatzas, and CV Jawahar. Reading between the lanes: Text videoqa on the road. In *International Conference on Document Analysis and Recognition*, pp. 137–154. Springer, 2023.

SMTI Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6, 2024.

Sameep Vani, Shreyas Jena, Maitreya Patel, Chitta Baral, Somak Aditya, and Yezhou Yang. Harnessing synthetic preference data for enhancing temporal understanding of video-llms. *ArXiv*, abs/2510.03955, 2025.

Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *ArXiv*, abs/2410.03290, 2024a.

Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290*, 2024b.

Hang Wang, Yuxiang Nie, Yongjie Ye, GuanYu Deng, Yanjie Wang, Shuai Li, Haiyang Yu, Jinghui Lu, and Can Huang. Dynamic-vlm: Simple dynamic visual token compression for videollm. *ArXiv*, abs/2412.09530, 2024c.

Haochen Wang, Qirui Chen, Cilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, Weidi Xie, and Stratis Gavves. Object-centric video question answering with visual grounding and referring. *ArXiv*, abs/2507.19599, 2025a.

Haochen Wang, Qirui Chen, Cilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, Weidi Xie, and Stratis Gavves. Object-centric video question answering with visual grounding and referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22274–22284, 2025b.

Haonan Wang, Hongfu Liu, Xiangyan Liu, Chao Du, Kenji Kawaguchi, Ye Wang, and Tianyu Pang. Fostering video reasoning via next-event prediction. *ArXiv*, abs/2505.22457, 2025c.

Jiankang Wang, Zhihan Zhang, Zhihang Liu, Yang Li, Jiannan Ge, Hongtao Xie, and Yongdong Zhang. Spacevllm: Endowing multimodal large language model with spatio-temporal video grounding capability. *arXiv preprint arXiv:2503.13983*, 2025d.

Jiawei Wang, Liping Yuan, and Yuchen Zhang. Tarsier: Recipes for training and evaluating large video description models. *ArXiv*, abs/2407.00634, 2024d.

Lan Wang, Yujia Chen, Du Tran, V. Boddeti, and Wen-Sheng Chu. Seal: Semantic attention learning for long video representation. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26192–26201, 2024e.

Mengzhao Wang, Huafeng Li, Yafei Zhang, Jinxing Li, Minghong Xie, and Dapeng Tao. Dual-task mutual reinforcing embedded joint video paragraph retrieval and grounding. *IEEE Transactions on Multimedia*, 27:5879–5894, 2024f.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *ArXiv*, abs/2505.12434, 2025e.

Qiang Wang, Xinyuan Gao, Songlin Dong, Jizhou Han, Jiangyang Li, Yuhang He, and Yihong Gong. Vdc-agent: When video detailed captioners evolve themselves via agentic self-reflection. *ArXiv*, abs/2511.19436, 2025f.

Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Vamos: Versatile action models for video understanding. *ArXiv*, abs/2311.13627, 2023a.

Shuai Wang, Daoan Zhang, Tianyi Bai, Shitong Shao, Jiebo Luo, and Jiaheng Wei. Last: Learning to think in space and time for generalist vision-language models. *ArXiv*, abs/2511.19261, 2025g.

Teng Wang, Jinrui Zhang, Feng Zheng, Wenhao Jiang, Ran Cheng, and Ping Luo. Learning grounded vision-language representation for versatile understanding in untrimmed videos. *ArXiv*, abs/2303.06378, 2023b.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025h.

Wentao Wang, Heqing Zou, Tianze Luo, Rui Huang, Yutian Zhao, Zhuochen Wang, Hansheng Zhang, Chengwei Qin, Yan Wang, Lin Zhao, and Huaijian Zhang. Video-str: Reinforcing mllms in video spatio-temporal reasoning with relation graph. *ArXiv*, abs/2510.10976, 2025i.

Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. Retake: Reducing temporal and knowledge redundancy for long video understanding. *ArXiv*, abs/2412.20504, 2024g.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and S. Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. *ArXiv*, abs/2403.10517, 2024h.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76. Springer, 2024i.

Xizi Wang, Feng Cheng, Ziyang Wang, Huiyu Wang, Md Mohaiminul Islam, Lorenzo Torresani, Mohit Bansal, Gedas Bertasius, and David Crandall. Timerefine: Temporal grounding with time refining video llm. *ArXiv*, abs/2412.09601, 2024j.

Yanan Wang, Julio Vizcarra, Zhi Li, Hao Niu, and Mori Kurokawa. Cotasks: Chain-of-thought based video instruction tuning tasks. *ArXiv*, abs/2507.13609, 2025j.

Ye Wang, Boshen Xu, Zihao Yue, Zihan Xiao, Ziheng Wang, Liang Zhang, Dingyi Yang, Wenxuan Wang, and Qin Jin. Time-r1: Post-training large vision language model for temporal video grounding. 2025k.

Ye Wang, Boshen Xu, Zihao Yue, Zihan Xiao, Ziheng Wang, Liang Zhang, Dingyi Yang, Wenxuan Wang, and Qin Jin. Timezero: Temporal video grounding with reasoning-guided lvlm. *ArXiv*, abs/2503.13377, 2025l.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding. pp. 396–416, 2024k.

Yicheng Wang, Zhikang Zhang, Jue Wang, D. Fan, Zhenlin Xu, Linda Liu, Xiang Hao, Vimal Bhat, and Xinyu Li. Gexia: Granularity expansion and iterative approximation for scalable multi-grained video-language learning. *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4725–4735, 2024l.

Yixu Wang, Jiaxin Song, Yifeng Gao, Xin Wang, Yang Yao, Yan Teng, Xingjun Ma, Yingchun Wang, and Yu-Gang Jiang. Safevid: Toward safety aligned video large multimodal models. *arXiv preprint arXiv:2505.11926*, 2025m.

Yiyu Wang, Xuyang Liu, Xiyan Gui, Xinying Lin, Boxue Yang, Chenfei Liao, Tailai Chen, and Linfeng Zhang. Accelerating streaming video large language models via hierarchical token compression. *ArXiv*, abs/2512.00891, 2025n.

Youze Wang, Zijun Chen, Ruoyu Chen, Shishen Gu, Yinpeng Dong, Hang Su, Jun Zhu, Meng Wang, Richang Hong, and Wenbo Hu. Understanding and benchmarking the trustworthiness in multimodal llms for video understanding. *arXiv preprint arXiv:2506.12336*, 2025o.

Yunxiao Wang, Mengxi Liu, Rui Shao, Haoyu Zhang, Bin Wen, Fan Yang, Tingting Gao, Di Zhang, and Liqiang Nie. Time: Temporal-sensitive multi-dimensional instruction tuning and benchmarking for video-llms. *ArXiv*, abs/2503.09994, 2025p.

Yuxuan Wang, Yiqi Song, Cihang Xie, Yang Liu, and Zilong Zheng. Videollamb: Long streaming video understanding with recurrent memory bridges. 2024m.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024n.

Yuxuan Wang, Cihang Xie, Yang Liu, and Zilong Zheng. Videollamb: Long-context video understanding with recurrent memory bridges. *ArXiv*, abs/2409.01071, 2024o.

Yuxuan Wang, Cihang Xie, Yang Liu, and Zilong Zheng. Videollamb: Long-context video understanding with recurrent memory bridges. *arXiv preprint arXiv:2409.01071*, 2024p.

Z. Wang, Fangtao Li, K. Ota, M. Dong, and Bin Wu. Regr: Relation-aware graph reasoning framework for video question answering. *Inf. Process. Manag.*, 60:103375, 2023c.

Zixuan Wang, Jinghao Shi, Hanzhong Liang, Xiang Shen, Vera Wen, Zhiqian Chen, Yifan Wu, Zhixin Zhang, and Hongyu Xiong. Filter-and-refine: A mllm based cascade system for industrial-scale video content moderation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 873–880, 2025q.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S. Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding. *ArXiv*, abs/2512.05774, 2025r.

Syed Talal Wasim, Muzammal Naseer, Salman H. Khan, Ming-Hsuan Yang, and F. Khan. Videogrounding-dino: Towards open-vocabulary spatio- temporal video grounding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18909–18918, 2023.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Yushen Wei, Yang Liu, Hong Yan, Guanbin Li, and Liang Lin. Visual causal scene refinement for video question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 377–386, 2023a.

Yushen Wei, Yang Liu, Hongfei Yan, Guanbin Li, and Liang Lin. Visual causal scene refinement for video question answering. *Proceedings of the 31st ACM International Conference on Multimedia*, 2023b.

Yixuan Weng and Bin Li. Visual answer localization with cross-modal mutual knowledge transfer. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2022.

Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. pp. 453–470, 2024.

Chang-Hsun Wu, Kai-Po Chang, Yu-Yang Sheng, Hung-Kai Chung, Kuei-Chun Wang, and Yu-Chiang Frank Wang. Season: Mitigating temporal hallucination in video large language models via self-diagnostic contrastive decoding. *arXiv preprint arXiv:2512.04643*, 2025a.

Hao Wu, Huabin Liu, Yu Qiao, and Xiao Sun. Dibs: Enhancing dense video captioning with unlabeled videos via pseudo boundary enrichment and online refinement. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18699–18708, 2024a.

Peiran Wu, Yunze Liu, Miao Liu, and Junxiao Shen. St-think: How multimodal large language models reason about 4d worlds from ego-centric videos. *arXiv preprint arXiv:2503.12542*, 2025b.

Shiwei Wu, Joya Chen, Kevin Qinghong Lin, Qimeng Wang, Yan Gao, Qianli Xu, Tong Xu, Yao Hu, Enhong Chen, and Mike Zheng Shou. Videollm-mod: Efficient video-language streaming with mixture-of-depths vision computation. *ArXiv*, abs/2408.16730, 2024b.

Tao Wu, Li Yang, Gen Zhan, Yabin Zhang, Yiting Liao, Junli Li, Deliang Fu, Li Zhang, and Limin Wang. Tempr1: Improving temporal understanding of mllms via temporal-aware multi-task reinforcement learning. *ArXiv*, abs/2512.03963, 2025c.

Xiyang Wu, Zongxia Li, Jihui Jin, Guangyao Shi, KV Gouthaman, Vishnu Raj, Nilotpal Sinha, Jingxi Chen, Fan Du, and Dinesh Manocha. Mass: Motion-aware spatial-temporal grounding for physics reasoning and comprehension in vision-language models. *ArXiv*, abs/2511.18373, 2025d.

Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13754–13765, 2025e.

Shao-Jun Xia, Huixin Zhang, Yifan Jiang, Xiaoyang Chen, Yike Chen, Zitong Li, Zhongwei Wan, and Anlan Sun. Rethinking hallucinations: A cognitive-inspired taxonomy and comprehensive survey in large language models, large vision-language models, and multimodal large language models. *Large Vision-Language Models, and Multimodal Large Language Models (November 09, 2025)*, 2025.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

Junbin Xiao, Pan Zhou, Tat seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. pp. 39–58, 2022.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13204–13214, 2023a.

Junbin Xiao, Pan Zhou, Angela Yao, Yicong Li, Richang Hong, Shuicheng Yan, and Tat seng Chua. Contrastive video question answering via video graph transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:13265–13280, 2023b.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13204–13214, 2024.

Jun Xie, Zhaoran Zhao, Xiongjun Guan, Yingjian Zhu, Hongzhu Yi, Xinming Wang, Feng Chen, and Zhepeng Wang. Team of one: Cracking complex video qa with model synergy. *ArXiv*, abs/2507.13820, 2025.

Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. *arXiv preprint arXiv:2501.13468*, 2025.

D. Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. *Proceedings of the 25th ACM international conference on Multimedia*, 2017. URL `https://api.semanticscholar.org/CorpusID:3864050`.

Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Slot-vlm: Slowfast slots for video-language modeling. *ArXiv*, abs/2402.13088, 2024a.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 2016. URL `https://www.microsoft.com/en-us/research/publication/msr-vtt-a-large-video-description-dataset-for-bridging-video-and-language/`.

Li Xu, He Huang, and Jun Liu. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9878–9888, 2021.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. Streamingvlm: Real-time understanding for infinite video streams. *ArXiv*, abs/2510.09608, 2025a.

Wenhao Xu, Xin Dong, Yue Li, Haoyuan Shi, and Zhiwei Xiong. Eventstu: Event-guided efficient spatio-temporal understanding for video large language models. *ArXiv*, abs/2511.18920, 2025b.

Yifan Xu, Xinhao Li, Yichun Yang, Desen Meng, Rui Huang, and Limin Wang. Carebench: A fine-grained benchmark for video captioning and retrieval. *arXiv preprint arXiv:2501.00513*, 2024b.

ZeYu Xu, Junkang Zhang, Qiang Wang, and Yi Liu. E-vrag: Enhancing long video understanding with resource-efficient retrieval augmented generation. *ArXiv*, abs/2508.01546, 2025c.

Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. *ArXiv*, abs/2408.10188, 2024.

Zhucun Xue, Jiang-She Zhang, Xurong Xie, Yuxuan Cai, Yong Liu, Xiangtai Li, and Dacheng Tao. Adavideorag: Omni-contextual adaptive retrieval-augmented efficient long video understanding. *ArXiv*, abs/2506.13589, 2025a.

Zihui Xue, Romy Mi Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. *ArXiv*, abs/2506.03340, 2025b.

Shilin Yan, Jiaming Han, Joey Tsai, Hongwei Xue, Rongyao Fang, Lingyi Hong, Ziyu Guo, and Ray Zhang. Crosslmm: Decoupling long video sequences from lmms via dual cross-attention mechanisms. *ArXiv*, abs/2505.17020, 2025.

Antoine Yang, Arsha Nagrani, P. H. Seo, Antoine Miech, J. Pont-Tuset, I. Laptev, Josef Sivic, and C. Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10714–10726, 2023.

Garry Yang, Zizhe Chen, Man Hon Wong, Haoyu Lei, Yongqiang Chen, Zhenguo Li, Kaiwen Zhou, and James Cheng. Mesh-understanding videos like human: Measuring hallucinations in large video models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4827–4836, 2025a.

Haolin Yang, Feilong Tang, Linxiao Zhao, Xiang An, Ming Hu, Huifa Li, Xinlin Zhuang, Boqian Wang, Yifan Lu, Xiaofeng Zhang, Abdalla Swikir, Junjun He, Zongyuan Ge, and Imran Razzak. Streamagent: Towards anticipatory agents for streaming video understanding. *ArXiv*, abs/2508.01875, 2025b.

Xun Yang, Jianming Zeng, Dan Guo, Shanshan Wang, Jianfeng Dong, and Meng Wang. Robust video question answering via contrastive cross-modality representation learning. *Science China Information Sciences*, 67, 2024.

Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, S. Mishra, Lizhu Zhang, and Mengye Ren. Streammem: Query-agnostic kv cache memory for streaming video understanding. *ArXiv*, abs/2508.15717, 2025c.

Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, Shlok Kumar Mishra, Lizhu Zhang, and Mengye Ren. Streammem: Query-agnostic kv cache memory for streaming video understanding. *arXiv preprint arXiv:2508.15717*, 2025d.

Yifeng Yao, Yike Yun, Jing Wang, Huishuai Zhang, Dongyang Zhao, Ke Tian, Zhihao Wang, Min Qiu, and Tao Wang. K-frames: Scene-driven any-k keyframe selection for long video understanding. *ArXiv*, abs/2510.13891, 2025.

Cheng Ye, Weidong Chen, Jingyu Li, Lei Zhang, and Zhendong Mao. Dual-path collaborative generation network for emotional video captioning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 496–505, 2024.

Jeong Hun Yeo, Sangyun Chung, Sungjune Park, Dae Hoe Kim, Jinyoung Moon, and Y. Ro. Gcagent: Long-video understanding via schematic and narrative episodic memory. *ArXiv*, abs/2511.12027, 2025.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Yunhang Shen, Chunjiang Ge, Yan Yang, Zuwei Long, Yuhan Dai, Tong Xu, Xing Sun, Ran He, Caifeng Shan, and Enhong Chen. Sparrow: Data-efficient video-llm with text-to-image augmentation. 2024.

Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. Videoarm: Agentic reasoning over hierarchical memory for long-form video understanding. *ArXiv*, abs/2512.12360, 2025.

Rongyi Yu, Chenyuan Duan, and Wentao Zhang. Longvidsearch: An agentic benchmark for multi-hop evidence retrieval planning in long videos. *arXiv preprint arXiv:2603.14468*, 2026.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *ArXiv*, abs/2305.06988, 2023.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-voyager: Learning to query frames for video large language models. *arXiv preprint arXiv:2410.03226*, 2024a.

Ting Yu, Kunhao Fu, Shuhui Wang, Qingming Huang, and Jun Yu. Prompting video-language foundation models with domain-specific fine-grained heuristics for video question answering. *IEEE Transactions on Circuits and Systems for Video Technology*, 35:1615–1630, 2024b.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 9127–9134, 2019.

Huaying Yuan, Zheng Liu, Minghao Qin, Hongjin Qian, Yan Shu, Zhicheng Dou, Ji-Rong Wen, and N. Sebe. Memory-enhanced retrieval augmentation for long video understanding. *ArXiv*, abs/2503.09149, 2025a.

Huaying Yuan, Zheng Liu, Minghao Qin, Hongjin Qian, Yan Shu, Zhicheng Dou, Ji-Rong Wen, and Nicu Sebe. Memory-enhanced retrieval augmentation for long video understanding. *arXiv preprint arXiv:2503.09149*, 2025b.

Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, Jianke Zhu, and Li Bing. Videorefer suite: Advancing spatial-temporal object understanding with video llm. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18970–18980, 2024.

Yunzhuzhang, Yu Lu, Tianyi Wang, Fengyun Rao, Yi Yang, and Linchao Zhu. Flexselect: Flexible token selection for efficient long video understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=0D3ja9s17M`.

Nianbo Zeng, Haowen Hou, F. Yu, Si Shi, and Y. He. Scenerag: Scene-level retrieval-augmented generation for video understanding. *ArXiv*, abs/2506.07600, 2025a.

Wenzheng Zeng, Difei Gao, Mike Zheng Shou, and Hwee Tou Ng. Factorized learning for temporally grounded video-language models. *ArXiv*, abs/2512.24097, 2025b.

Yingsen Zeng, Zepeng Huang, Yujie Zhong, Chengjian Feng, Jie Hu, Lin Ma, and Yang Liu. Distime: Distribution-based time representation for video large language models. *ArXiv*, abs/2505.24329, 2025c.

Ziyun Zeng, Yixiao Ge, Zhan Tong, Xihui Liu, Shutao Xia, and Ying Shan. Tvtsv2: Learning out-of-the-box spatiotemporal visual representations at scale. *ArXiv*, abs/2305.14173, 2023.

Yu-Wei Zhan, Xin Wang, Hong Chen, Tongtong Feng, Wei Feng, Ren Wang, Guangyao Li, Qing Li, and Wenwu Zhu. Phyvllm: Physics-guided video language model with motion-appearance disentanglement. *ArXiv*, abs/2512.04532, 2025a.

Yu-Wei Zhan, Xin Wang, Hong Chen, Tongtong Feng, Wei Feng, Ren Wang, Guangyao Li, Qing Li, and Wenwu Zhu. Phyvllm: Physics-guided video language model with motion-appearance disentanglement. *arXiv preprint arXiv:2512.04532*, 2025b.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. pp. 21715–21737, 2023a.

Congzhi Zhang, Zhibin Wang, Yinchao Ma, Jiawei Peng, Yihan Wang, Qiang Zhou, Jun Song, and Bo Zheng. Rewatch-r1: Boosting complex video reasoning in large vision-language models through agentic data synthesis. *ArXiv*, abs/2509.23652, 2025a.

Deyu Zhang, Tingting Long, Jinrui Zhang, Ligeng Chen, Ju Ren, and Yaoxue Zhang. Prompt-aware of frame sampling for efficient text-video retrieval. *arXiv preprint arXiv:2507.15491*, 2025b.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. pp. 543–553, 2023b.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *ArXiv*, abs/2406.08085, 2024a.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. *ArXiv*, abs/2508.04416, 2025c.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-vstream: Efficient real-time understanding for long video streams. *ArXiv*, abs/2506.23825, 2025d.

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Jingjing Chen, and Yu-Gang Jiang. Eventhallusion: Diagnosing event hallucinations in video llms. *ArXiv*, abs/2409.16597, 2024b.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025e.

Xian Zhang, Zexi Wu, Zinuo Li, Hongming Xu, Luqi Gong, Farid Boussaid, Naoufel Werghi, and Mohammed Bennamoun. Adard-key: Adaptive relevance-diversity keyframe sampling for long-form video understanding. *arXiv preprint arXiv:2510.02778*, 2025f.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *ArXiv*, abs/2505.18079, 2025g.

Yan Zhang, Gangyan Zeng, Daiqing Wu, Huawen Shen, Binbin Li, Yu ZHOU, Can Ma, and Xiaojun Bi. *Gather and Trace: Rethinking Video TextVQA from an Instance-oriented Perspective*. 2025h.

Yongheng Zhang, Xu Liu, Ruihan Tao, Qiguang Chen, Hao Fei, Wanxiang Che, and Libo Qin. Vitcot: Video-text interleaved chain-of-thought for boosting video understanding in large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025i.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *Trans. Mach. Learn. Res.*, 2025, 2024c.

Yue Zhang, Jilei Sun, Yunhui Guo, and Vibhav Gogate. Can video large multimodal models think like doubters-or double-down: A study on defeasible video entailment. *ArXiv*, abs/2506.22385, 2025j.

Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10668–10677, 2020.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

Fufangchen Zhao, Liao Zhang, Daiqi Shi, Yuanjun Gao, Chen Ye, Yang Cai, Jian Gao, and Danfeng Yan. Videoperceiver: Enhancing fine-grained temporal perception in video multimodal large language models. *ArXiv*, abs/2511.18823, 2025a.

Lingjun Zhao, Mingyang Xie, Paola Cascante-Bonilla, Hal Daum'e, and Kwonjoon Lee. Can hallucination correction improve video-language alignment? *ArXiv*, abs/2502.15079, 2025b.

Zijia Zhao, Yuqi Huo, Tongtian Yue, Longteng Guo, Haoyu Lu, Bingning Wang, Weipeng Chen, and Jing Liu. Efficient motion-aware video mllm. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24159–24168, 2025c.

Zhuo Zhi, Qiangqiang Wu, Minghe Shen, Wenbo Li, Yinchuan Li, Kun Shao, and Kaiwen Zhou. Videoagent2: Enhancing the llm-based agent system for long-form video understanding by uncertainty-aware cot. *ArXiv*, abs/2504.04471, 2025.

Chunlin Zhong, Qiuxia Hou, Zhangjun Zhou, Shuang Hao, Haonan Lu, Yanhao Zhang, He Tang, and Xiang Bai. Owlcap: Harmonizing motion-detail for video captioning via hmd-270k and caption set equivalence reward. *ArXiv*, abs/2508.18634, 2025.

Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 6439–6455, 2022.

Honglu Zhou, Xiangyu Peng, Shrikant Kendre, Michael S. Ryoo, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. Strefer: Empowering video llms with space-time referring and reasoning via synthetic instruction data. *ArXiv*, abs/2509.03501, 2025a.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025b.

Pengyuan Zhou, Lin Wang, Zhi Liu, Yanbin Hao, Pan Hui, Sasu Tarkoma, and Jussi Kangasharju. A survey on generative ai and llm for video generation, understanding, and streaming. *arXiv preprint arXiv:2404.16038*, 2024.

Shengchao Zhou, Yuxin Chen, Yuying Ge, Wei Huang, Jiehong Lin, Ying Shan, and Xiaojuan Qi. Learning to reason in 4d: Dynamic spatial understanding for vision language models. *ArXiv*, abs/2512.20557, 2025c.

Yiyang Zhou, Yangfan He, Yaofeng Su, Siwei Han, Joel Jang, Gedas Bertasius, Mohit Bansal, and Huaxiu Yao. Reagent-v: A reward-driven multi-agent framework for video understanding. *ArXiv*, abs/2506.01300, 2025d.

Yuan Zhou, Litao Hua, Shilong Jin, Wentao Huang, and Haoran Duan. Reason: Reinforced causal search with information bottleneck for video understanding. *ArXiv*, abs/2511.12530, 2025e.

Wencheng Zhu, Yuexin Wang, Hongxuan Li, Pengfei Zhu, and Qinghua Hu. Vtd-clip: Video-to-text discretization via prompting clip. *ArXiv*, abs/2503.18407, 2025.

Orr Zohar, Xiaohan Wang, Yonatan Bitton, Idan Szpektor, and S. Yeung-Levy. Video-star: Self-training enables video instruction tuning with any supervision. *ArXiv*, abs/2407.06189, 2024.

Heqing Zou, Tianze Luo, Guiyang Xie, Fengmao Lv, Guangcong Wang, Junyang Chen, Zhuochen Wang, Hansheng Zhang, Huaijian Zhang, et al. From seconds to hours: Reviewing multimodal large language models on comprehensive long video understanding. *arXiv preprint arXiv:2409.18938*, 2024.

Jialong Zuo, Yongtai Deng, Lingdong Kong, Jingkang Yang, Rui Jin, Yiwei Zhang, Nong Sang, Liang Pan, Ziwei Liu, and Changxin Gao. Videolucy: Deep memory backtracking for long video understanding. *ArXiv*, abs/2510.12422, 2025.