Distorted or Fabricated? A Survey on Hallucination in Video LLMs

Anonymous ACL submission

Abstract

Despite significant progress in video-language modeling, hallucinations remain a persistent challenge in Video Large Language Models (Vid-LLMs), referring to outputs that appear plausible yet contradict the content of the input video. This survey presents a comprehensive analysis of hallucinations in Vid-LLMs and introduces a systematic taxonomy that categorizes them into two core types: dynamic distortion and content fabrication, each comprising two subtypes with representative cases. Building on this taxonomy, we review recent advances in the evaluation and mitigation of hallucinations, covering key benchmarks, metrics, and intervention strategies. We further analyze the root causes of dynamic distortion and content fabrication, which often result from limited capacity for temporal representation and insufficient visual grounding. These insights inform several promising directions for future work, including the development of motion-aware visual encoders and the integration of counterfactual learning techniques. This survey consolidates scattered progress to foster a systematic understanding of hallucinations in Vid-LLMs, laying the groundwork for building robust and reliable video-language systems.

1 Introduction

Video Large Language Models (Vid-LLMs) extend the capabilities of vision-language systems from static images to temporally coherent video inputs, enabling tasks such as action recognition, temporal reasoning, and audio-visual understanding (Zhang et al., 2023a; Maaz et al., 2024; Li et al., 2023a; Lin et al., 2024; Wang et al., 2024a; Fu et al., 2024). Despite recent advances, these models remain susceptible to hallucinations, producing outputs that appear plausible and coherent yet contradict the actual content of the video. This issue poses reliability and safety risks in safety-critical domains, including embodied AI (Wu et al., 2023) and autonomous driving (Chen et al., 2024).

While hallucinations have been extensively surveyed in image-based vision-language models (VLMs) (Liu et al., 2024a; Lan et al., 2024), the inherent complexity of video's temporal structure, motion dynamics, and audio-visual integration complicates the direct application of these insights to the video domain. To address this gap, this survey presents a video-specific, mechanism-driven taxonomy that classifies hallucinations into two primary types: dynamic distortion, where the model misrepresents the spatiotemporal evolution or referential consistency of entities and scenes; and content fabrication, where outputs are influenced by prior knowledge or dominated by audio modality.

Building on this taxonomy, we review recent advances in the evaluation and mitigation of hallucinations in Vid-LLMs, with a focus on key benchmarks, metrics, and intervention strategies. Dynamic distortion includes hallucinations in spatiotemporal dynamics, such as incorrect event ordering (Li et al., 2025a; Wu et al., 2025a; Sun et al., 2025), inaccurate duration estimation (Wang et al., 2024b; Huang et al., 2025d; Sun et al., 2024), and frequency miscounting (Gao et al., 2025; Choong et al., 2024), as well as referential inconsistency, where the model conflates different characters (Seth et al., 2025; Yang et al., 2025) or scenes (Lu et al., 2025; Ma et al., 2024; Pu et al., 2025). Content fabrication includes context-driven hallucinations, where commonly co-occurring object–action (Chang et al., 2025; Li et al., 2025c) or scene–event (Bae et al., 2025; Zhang et al., 2024; Ding et al., 2025) patterns lead to unsupported inferences; and audio-visual conflict, where dominant auditory cues override visual evidence, resulting in hallucinated actions (Sung-Bin et al., 2025; Jung et al., 2025) or emotional states (Xing et al., 2025).

Further analysis reveals the underlying mechanisms of dynamic distortion and content fabrication. Dynamic distortion often results from missing fine-grained motion cues due to limited temporal

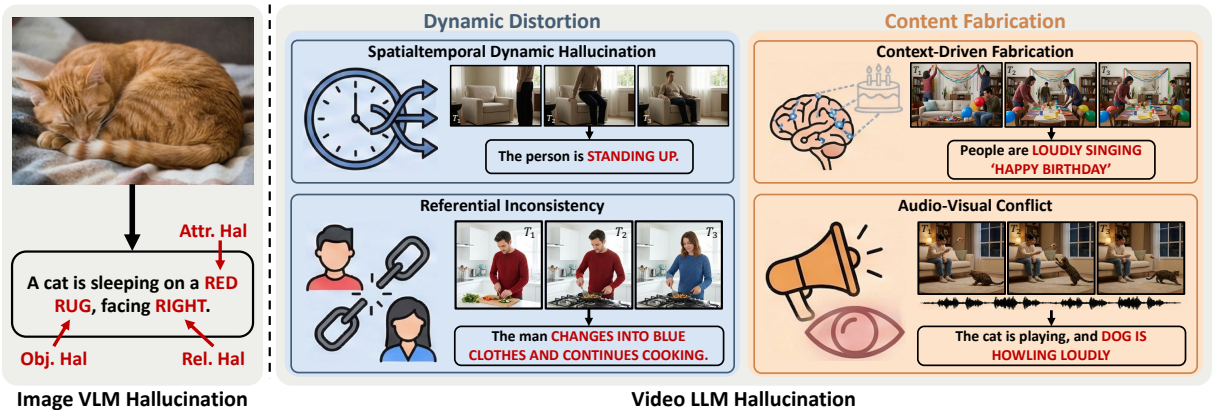

Figure 1: Taxonomy of hallucinations in Vid-LLMs. Video models exhibit unique hallucination types beyond static errors in images. These fall into two main categories: (1) Dynamic Distortion, which includes spatiotemporal misrepresentation and referential inconsistency; (2) Content Fabrication, which includes hallucinations influenced by statistical priors and cases where auditory input overrides visual evidence.

encoding (Zhao et al., 2025; Liu et al., 2024b), and is further exacerbated in long videos by weak long-range memory (Bae et al., 2025) and poor temporal localization (Wu et al., 2025a). In contrast, content fabrication arises from insufficient visual grounding (Lee et al., 2025), allowing pretrained priors (Li et al., 2025c) or dominant audio signals (Leng et al., 2024) to override visual evidence.

In light of these underlying mechanisms, promising research directions include developing motion-aware architectures(Wu et al., 2025b) that retain fine-grained temporal features to strengthen the alignment between visual perception and temporal reasoning. In addition, counterfactual training strategies that disentangle visual evidence from prior knowledge(Huang et al., 2025c) offer a principled approach to mitigating content fabrication by encouraging models to ground predictions more faithfully in the visual input.

Comparison with existing surveys. Hallucination has been extensively studied in LLMs and image-based VLMs (Zhang et al., 2023c; Huang et al., 2025b; Liu et al., 2024a; Lan et al., 2024). While MLLM hallucination surveys (Sahoo et al., 2024; Bai et al., 2024) include video alongside other modalities, their discussion of video hallucination remains superficial, offering only brief mentions of benchmarks and mitigation strategies without structural or causal analysis. In contrast, this survey presents the first mechanism-driven taxonomy of hallucinations in Vid-LLMs. We propose a layered classification framework (Fig. 2), conduct a broader and more detailed review of existing literature, and analyze the underlying causes of hallucinations. Building on this analysis, we outline future directions that align closely with identified causes, benchmark coverage, and mitigation strategies, offering a cohesive roadmap toward hallucination-resilient Vid-LLMs.

2 Definition and Scope

Definition. We define *video hallucination* as cases where a Vid-LLM generates textual outputs that are linguistically coherent and contextually plausible, yet contradict the observable spatiotemporal evidence in the input video.

Distinction from Static Image Hallucination. While hallucinations in image-based VLMs, such as those involving objects, attributes, and relations, have been well studied (Liu et al., 2024a; Lan et al., 2024), the video modality introduces a temporal dimension that fundamentally alters the problem. Unlike static inputs, videos require reasoning over causality, temporal grounding, motion dynamics, and audio-visual integration. These aspects are beyond the scope of traditional static metrics. Our taxonomy explicitly captures these temporal and multimodal challenges, distinguishing video hallucination from image-based settings.

3 Taxonomy of Video Hallucinations

As discussed in Section 2, the temporal and multimodal nature of video poses challenges beyond static image settings. To address these, we propose a mechanism-driven taxonomy focused on *dynamic-level hallucinations* unique to video. It captures how Vid-LLMs fail in temporal reasoning and cross-modal alignment. We classify hallucina-

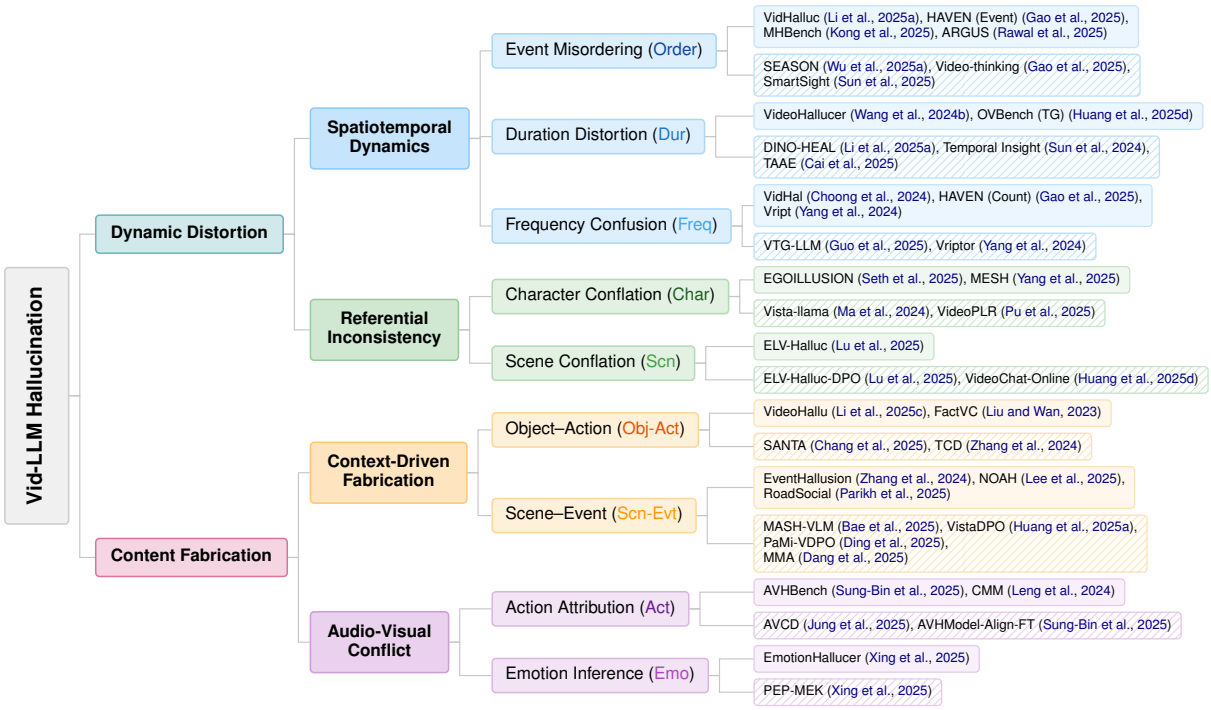

Figure 2: Mechanism-driven taxonomy of Vid-LLM hallucinations. **Dynamic Distortion**: entities are perceived but their spatiotemporal evolution or identity is misinterpreted, including Spatiotemporal Dynamics (Order/Dur/Freq) and Referential Inconsistency (Char/Scn). **Content Fabrication**: outputs lack visual evidence and are driven by priors, including Context-Driven Fabrication (Obj-Act/Scn-Evt) and Audio-Visual Conflict (Act/Emo). Solid fill denotes benchmarks; striped fill indicates mitigation methods.

tions into *Dynamic Distortion* and *Content Fabrication*, each with two subtypes and representative cases (Figure 2). This taxonomy provides a foundation for structuring benchmarks (Section 4) and mitigation strategies (Section 5).

3.1 Dynamic Distortion

This category refers to situations where the model correctly detects entities but misrepresents their temporal progression or referential consistency. It includes two subtypes: *Spatiotemporal Dynamics*, involving errors in event ordering, duration, or frequency; and *Referential Inconsistency*, where characters or scenes are conflated across temporal boundaries.

Spatiotemporal Dynamics. These hallucinations arise when the model correctly identifies relevant events but fails to model their temporal relationships. Typical cases include event misordering, such as reversing action causality or misinterpreting motion direction and trajectory (Li et al., 2025a; Gao et al., 2025; Wu et al., 2025a; Sun et al., 2025); duration distortion, where the model over- or underestimates the length of an action (Wang et al., 2024b; Huang et al., 2025d; Sun et al., 2024); and frequency confusion, in which repeated actions are miscounted (Gao et al., 2025; Choong et al., 2024).

Referential Inconsistency These hallucinations refers to semantic-level failures where the model conflates distinct entities or scenes across temporal boundaries, producing blended descriptions that obscure segment distinctions. These errors arise when content from separate time spans is incorrectly merged into a single entity- or scene-level statement, even when visual cues could distinguish them. Such inconsistency typically appears in two forms: character conflation, where different individuals across scenes are mistakenly treated as the same person (Seth et al., 2025; Yang et al., 2025); and scene conflation, where actions or settings from distinct contexts are combined into a single narrative (Lu et al., 2025; Pu et al., 2025).

3.2 Content Fabrication

This category covers cases where the model produces outputs that lack grounding in visual evidence and are instead influenced by learned priors. It includes *context-driven fabrication*, where common object–action or scene–event associations result in unsupported predictions, and *audio-visual conflict*, where auditory cues override visual input.

Context-Driven Fabrication This type of hallu-

cination arises when the model relies on statistical associations from training data rather than grounding its predictions in visual evidence. An error is considered context-driven fabrication when the predicted action or event lacks visual support in the current observation window but is triggered by the presence of associated objects or scenes. It typically appears in two forms: object–action fabrication and scene–event fabrication. Object–action fabrication (Chang et al., 2025; Li et al., 2025c; Liu and Wan, 2023) occurs when the presence of an object leads to incorrect action inference despite lacking motion cues. Scene–event fabrication (Bae et al., 2025; Zhang et al., 2024; Ding et al., 2025; Dang et al., 2025; Huang et al., 2025a) happens when typical events are predicted solely from background settings.

Audio-Visual Conflict This type of fabrication occurs when dominant or misleading audio cues override visual evidence, leading the model to generate outputs that align more with the audio than the video. Typical cases include hallucinated actions triggered by background sounds (Sung-Bin et al., 2025; Jung et al., 2025), and emotion inference based on vocal tone rather than facial expression (Xing et al., 2025).

4 Evaluation Benchmarks

Following the taxonomy in Section 3, we categorize existing benchmarks by hallucination type and representative failure cases. Table 1 provides an overview of their venues, scales, task formats, and evaluation metrics.

4.1 Dynamic Distortion Benchmarks

Spatiotemporal dynamics benchmarks assess Vid-LLMs' ability to model temporal structure, covering three subtypes: event misordering, duration distortion, and frequency confusion.

For *event misordering*, VidHalluc (Li et al., 2025a) includes 5,002 videos from ActivityNet (Yu et al., 2019), YouCook2 (Zhou et al., 2018), and VALOR32K (Liu et al., 2025a), and evaluates temporal hallucinations through sequence-based QA tasks that test whether models can determine the correct order of actions. HAVEN (Gao et al., 2025) (event) targets discrepancies in action sequences using 2,245 questions across binary, multiple-choice, and short-answer formats. MHBench (Kong et al., 2025) provides 1,200 videos and tests motion understanding via adversarial triplets simulating original, reversed, and incomplete actions. ARGUS (Rawal et al., 2025) evaluates hallucination and omission on 500 videos with about 9,500 annotations, penalizing event misordering by checking the temporal alignment between model-generated and ground-truth action sequences.

For *duration distortion*, VideoHallucer (Wang et al., 2024b) includes 1,800 adversarial question pairs based on 948 videos, assessing both intrinsic and extrinsic hallucinations through tasks focused on detecting abnormal durations and comparing relative event lengths. OVBench (THV) (Huang et al., 2025d) targets duration distortion in real-time streaming settings, requiring models to track action persistence and estimate the length of ongoing events as temporal context unfolds.

For *frequency confusion*, VidHal (Choong et al., 2024) benchmarks fine-grained temporal understanding by asking models to distinguish between captions with correct and hallucinated action counts. HAVEN (Gao et al., 2025) (count) addresses this via numerical questions that test a model's ability to differentiate between single and repeated actions, evaluating its capacity to quantify frequency. Vript (Yang et al., 2024) includes a 'Count' category in its Vript-RR benchmark, which assesses whether models can accurately compare the number of visual elements across long video sequences.

Referential inconsistency benchmarks assess whether models can maintain distinct representations of entities and scenes over time. Despite the increasing use of Vid-LLMs, only three benchmarks explicitly address this issue.

For *character conflation*, EGOILLUSION (Seth et al., 2025) includes 1,400 egocentric videos and 8,000 question–answer pairs. It evaluates whether models confuse different individuals, for example by identifying the camera wearer as another person during object interactions or activity recognition. MESH (Yang et al., 2025) introduces a human-aligned evaluation framework called Mise En Scène, built on TVQA+ clips. It tests whether models can consistently track character identity, appearance, and actions across scenes using structured evaluation traps.

For *scene conflation*, ELV-Halluc (Lu et al., 2025) uses approximately 8,600 adversarial video and text pairs to evaluate whether models incorrectly assign visual elements such as objects or actions from one part of a video to another.

Table 1: Summary of video hallucination benchmarks. *Format*: **MC** = Multiple Choice, **Bin** = Yes/No, **Open** = Open-ended QA, **Cap** = Captioning. *Length*: **S** = Short (<1min), **M** = Medium (1–5min), **L** = Long (>5min), **St** = Streaming. *Baseline*: Specialized baseline method proposed. *SOTA Perf.*: Representative best performance reported.

Benchmark	Venue	# Vid	# QA	Format	Metric	Len	Domain	Baseline	SOTA Perf.
Spatiotemporal Dynamics Benchmarks (Dynamic Distortion)									
VidHalluc (Li et al., 2025a)	CVPR'25	5,002	9,295	MC, Bin	Acc, Score	S	ActivityNet, YouCook2, VALOR	✓	GPT-4o: 81.2%
VideoHallucer (Wang et al., 2024b)	arXiv'24	948	1,800	Bin	Acc, Score	M	ActivityNet, VidOR, YouCook	✓	Gemini-1.5: 37.8%
HAVEN (Gao et al., 2025)	arXiv'25	–	6.5k	MC, Bin, Open	Acc, Bias	S	COIN, ActivityNet, Sports1M	✓	Valley-Eagle: 61.3%
MHBench (Kong et al., 2025)	AAAI'25	1,200	–	MC, Bin	Acc, F1	S	Sth-Sth V2, Self-shot	✓	VideoChat2-MCD: 65.2%
VidHal (Choong et al., 2024)	arXiv'24	1,000	3,000	MC	Acc, NDCG	S	TempCompass, MVBench, PT	✗	GPT-4o: 77.2%
ARGUS (Rawal et al., 2025)	arXiv'25	500	∼9.5k	Cap	Cost-H/O	S	Ego4D, Panda-70M, Stock	✗	Gemini-2.0: 41%
OVBench (THV) (Huang et al., 2025d)	CVPR'25	–	∼33k	Bin	Accuracy	St	DiDeMo, QuerYD.	✓	VideoChat-On: 63.1%
Vript (Yang et al., 2024)	NeurIPS'24	12k	420k	Bin, MC	Acc, F1	L	HD-VILA, YouTube, TikTok	✓	Vriptor: 58.3 (F1)
Referential Inconsistency Benchmarks (Dynamic Distortion)									
EGOILLUSION (Seth et al., 2025)	EMNLP'25	1,400	8k	Bin, Open	Accuracy	S/M	Ego4D, EgoSeg, Trek-150	✗	Gemini-Pro: 59.4%
MESH (Yang et al., 2025)	MM'25	–	∼140k	MC, Bin	Accuracy	S/M	TVQA+, UCF101	✗	GPT-4o: 79.1%
ELV-Halluc (Lu et al., 2025)	arXiv'25	200	4.8k	Bin	Acc, SAH	L	YouTube (Event-based)	✓	Gemini2.5-Flash: 53.1%
Context-Driven Fabrication Benchmarks (Content Fabrication)									
FactVC (Liu and Wan, 2023)	EMNLP'23	300	–	Cap	Bleu4, Rouge-L	M/L	ActivityNet, YouCook2	✓	PDVC-gt: 12.83 (Bleu4)
EventHallusion (Zhang et al., 2024)	AAAI'26	400	711	Bin, Open	Accuracy	S	ActivityNet	✓	GPT-4o: 91.93%
NOAH (Lee et al., 2025)	arXiv'25	9k	∼60k	Bin, Cap	Acc, HR	M/L	ActivityNet	✗	Gemini2.5-Flash: 66.8%
VideoHallu (Li et al., 2025c)	NeurIPS'25	987	3,233	Open	GPT-Score	S	Generated (Sora, etc.)	✓	Comb-GRPO: 57.7
RoadSocial (Parikh et al., 2025)	CVPR'25	13.2k	260k	Open	GPT-Score	S/M	Social Media (Traffic)	✗	GPT-4o: 69.8
Audio-Visual Conflict Benchmarks (Content Fabrication)									
AVHBench (Sung-Bin et al., 2025)	ICLR'25	2,136	5.3k	Bin	Accuracy	S	AudioCaps, VALOR	✓	AVHModel-Align-FT: 83.9%
CMM (Leng et al., 2024)	arXiv'24	1.2k	2.4k	Bin	PA/HR	S	WebVid, AudioCaps	✗	Gemini-1.5: 88.4/64.2
EmotionHallucer (Xing et al., 2025)	arXiv'25	230	2,742	Bin	Accuracy	S/M	MER 2023, Social-IQ 2.0	✓	Gemini2.5-Flash: 68.2%

4.2 Content Fabrication Benchmarks

Context-driven fabrication benchmarks assess whether models generate outputs based on visual evidence rather than relying on statistical associations from training data. These benchmarks span various domains such as activity recognition, driving, and synthetic videos, reflecting the diverse and context-sensitive nature of fabrication errors.

For *object-action hallucination*, VideoHallu (Li et al., 2025c) uses synthetic "negative control" videos to test whether models incorrectly infer actions based on prior object–action associations instead of actual motion cues. For instance, a model may claim that a watermelon breaks after being shot even when it remains intact in the video. FactVC (Liu and Wan, 2023) identifies action consistency as a major source of captioning error, accounting for 38.3% of failures. Models often describe interactions such as a person dancing with a dog based on object co-occurrence, without grounding predictions in visual dynamics.

For *scene-event hallucination*, EventHallusion (Zhang et al., 2024) evaluates whether models hallucinate events by over-relying on typical scene–event pairings, such as assuming cooking takes place in a kitchen even without action evidence. NOAH (Lee et al., 2025) scales this evaluation to over 60,000 samples created from around 15,000 edited videos, testing whether models ignore inserted contradictory clips and instead generate events that align with the surrounding scene or narrative context. RoadSocial (Parikh et al., 2025) focuses on driving scenarios, using adversarial and incompatible question formats to test whether models hallucinate common road events, such as collisions or traffic violations, based solely on general road context or misleading prompts, even when no such events occur in the video.

Audio-visual conflict benchmarks evaluate whether models integrate audio and visual signals appropriately, focusing on cases where dominant audio cues override visual input and lead to incorrect predictions. With only three existing benchmarks, this category remains underexplored, and current datasets are limited to short video clips. As multimodal Vid-LLMs increasingly process audio, further benchmark development is needed.

For *action attribution*, AVHBench (Sung-Bin et al., 2025) tests whether sounds such as music or bird calls cause models to generate incorrect visual descriptions like "a person is dancing" or "a bird is chirping," even when no such actions are visible. It includes 2,136 videos and 5,302 binary question–answer pairs sourced from AudioCaps and VALOR, and reports precision, recall, and F1 scores to quantify errors. CMM (Leng et al., 2024) evaluates similar cases using curated "audio dominance" samples, where prominent sounds such as thunder occur without visual events like lightning. Models are asked binary questions to assess whether they mistakenly rely on audio alone for visual claims.

For *emotion inference*, EmotionHallucer (Xing et al., 2025) examines whether models infer incor-

rect emotional states based on misleading multimodal cues. Its Reasoning Result and Reasoning Cue tasks test if models describe a neutral face as "excited" due to upbeat vocal tone, or invent emotional cues to justify unsupported conclusions.

4.3 Discussion: Coverage and Gaps

Table 1 summarizes 19 existing benchmarks for evaluating video hallucination, with a notable concentration on Spatiotemporal Dynamics (8 benchmarks), mostly targeting short clips. A few, such as Vript (Yang et al., 2024) and OVBench (Huang et al., 2025d), extend to long-form or streaming contexts. Context-Driven Fabrication shows broad domain coverage, ranging from traffic scenarios (RoadSocial) to synthetic videos (VideoHallu). In contrast, Referential Inconsistency and Audio-Visual Conflict remain underexplored, each represented by only three benchmarks, and no benchmark addresses audio-visual consistency in long-form videos. While 11 benchmarks include dedicated baselines to support method development, performance analysis reveals a clear divide: state-of-the-art models perform well on some tasks (e.g., VidHalluc, EventHallusion, with scores above 80%), but struggle with fine-grained temporal reasoning (e.g., VideoHallucer, 37.8%) and long-context consistency (e.g., ELV-Halluc, 53.1%). These findings identify dynamic distortion and long-range temporal grounding as persistent challenges for future research.

5 Mitigation Strategies

Following the taxonomy in Section 3, we group existing mitigation strategies by hallucination type and representative failure cases. Table 2 summarizes the corresponding techniques for each case.

5.1 Mitigating Dynamic Distortion

Spatiotemporal dynamics mitigation tackles event order, duration, and frequency Hallucinations using contrastive, optimization-based, and temporal grounding strategies, with event misordering being the most extensively studied.

For *event misordering*, SEASON (Wu et al., 2025a) contrasts original videos with temporally homogenized negatives that disrupt causal order, using self-diagnostic decoding to suppress outputs insensitive to correct sequence. Video-thinking (Gao et al., 2025) introduces TDPO (Thinking-based DPO), applying segment-weighted preference learning on reasoning paths to optimize for temporal logic. SmartSight (Sun et al., 2025) ranks multiple responses based on the Temporal Attention Collapse (TAC) score, favoring outputs that attend proportionally across time to preserve correct order.

For *duration distortion*, Temporal Insight Enhancement (Sun et al., 2024) decomposes events into atomic actions and leverages external vision models to timestamp them, aligning model responses with grounded temporal claims. DINO-HEAL (Li et al., 2025a) uses DINOv2-guided spatial saliency to reweight features and maintain attention on action-relevant regions across time. TAAE (Cai et al., 2025) identifies activation offsets between full and downsampled inputs to amplify duration-sensitive representations during inference.

For *frequency confusion*, VTG-LLM (Guo et al., 2025) introduces absolute-time tokens that decouple event identity from repetition count, improving temporal anchoring of repeated actions. Vriptor (Yang et al., 2024) aligns dense scene-level captions with timestamps to enforce instance-level discrimination, helping models avoid merging or duplicating repeated events in long videos.

Referential inconsistency mitigation focuses on preserving distinct representations of entities and scenes across time.

For *character conflation*, Vista-LLaMA (Ma et al., 2024) introduces Equal Distance Attention, which removes positional decay between visual and textual tokens, ensuring stable attention to character identities regardless of when they appear. VideoPLR (Pu et al., 2025) constructs a structured video database with explicit object tracking, allowing symbolic logic programs to differentiate entity identities during reasoning.

For *scene conflation*, ELV-Halluc-DPO (Lu et al., 2025) applies adversarial preference optimization using cross-segment perturbations that swap entities in space or time, encouraging the model to ground predictions within the correct segment. VideoChat-Online (Huang et al., 2025d) uses a Pyramid Memory Bank during streaming inference to separate recent high-resolution frames from compressed long-term history, reducing confusion between distinct temporal contexts.

5.2 Mitigating Content Fabrication

Context-driven fabrication mitigation aims to separate predictions from statistical associations and strengthen visual grounding.

For *object–action hallucination*, methods em-

Table 2: Summary of video hallucination mitigation strategies. *Case*: Order/Dur/Freq (Spatiotemporal Dynamics), Char/Scn (Referential Inconsistency), Obj-Act/Scn-Evt (Context-Driven Fabrication), Act/Emo (Audio-Visual Conflict). *TF*: ✓ = train-free, ✗ = training required. *Reported Gain*: reported improvement (over baseline) on primary benchmark.

Method	Venue	Case	TF	Core Technique	Key Mechanism	Reported Gain
Spatiotemporal Dynamics Mitigation (Dynamic Distortion)						
SEASON (Wu et al., 2025a)	arXiv'25	Order	✓	Contrastive Decoding	Temporal homogenization contrast	+5.7% Acc (Qwen2.5-VL)
Video-thinking (Gao et al., 2025)	arXiv'25	Order	✗	Preference Optimization	Segment-weighted thinking contrast	+7.4% Acc (LLaVA-NeXT)
SmartSight (Sun et al., 2025)	arXiv'25	Order	✓	Introspective Sampling	Temporal attention collapse score	+2.9% Acc (Video-R1)
Temporal Insight (Sun et al., 2024)	ICPR'24	Dur	✓	Post-hoc Correction	Iconic action timestamp extraction	+27.9% R@1 (Video-LLaMA)
DINO-HEAL (Li et al., 2025a)	CVPR'25	Dur	✓	Feature Reweighting	DINOv2 spatial saliency reweighting	+7.0% Acc (Video-LLaVA)
TAAE (Cai et al., 2025)	arXiv'25	Dur	✗	Activation Engineering	Temporal-aware offset injection	+4.4% Acc (Qwen2.5-VL)
VTG-LLM (Guo et al., 2025)	AAAI'25	Freq	✗	Temporal Grounding	Absolute-time token disentanglement	+6.8% R@1 (Video-LLaMA2)
Vriptor (Yang et al., 2024)	NeurIPS'24	Freq	✗	Video-Script Alignment	Dense script-based timestamp training	+7.5% F1 (ST-LLM)
Referential Inconsistency Mitigation (Dynamic Distortion)						
Vista-llama (Ma et al., 2024)	CVPR'24	Char	✗	Token Processing	Equal distance visual attention	~5.0% Acc (LLaVA)
VideoPLR (Pu et al., 2025)	arXiv'25	Char	✗	Perception-Logic-Reasoning	Database-anchored symbolic execution	+9.2% Acc (Qwen2.5-VL)
ELV-Halluc-DPO (Lu et al., 2025)	arXiv'25	Scn	✗	Preference Optimization	Cross-segment adversarial DPO	-27.7% SAH (Qwen2.5-VL)
VideoChat-Online (Huang et al., 2025d)	CVPR'25	Scn	✗	Streaming Processing	Pyramid memory bank update	+8.5% Acc (InternVL2)
Context-Driven Fabrication Mitigation (Content Fabrication)						
SANTA (Chang et al., 2025)	arXiv'25	Obj-Act	✗	Fine-grained Contrastive Tuning	Hard negative action/object swapping	+2.4% Acc (LLaVA-Video)
TCD (Zhang et al., 2024)	AAAI'26	Obj-Act	✓	Contrastive Decoding	Logit subtraction of priors	+3.2% Acc (VILA)
MASH-VLM (Bae et al., 2025)	CVPR'25	Scn-Evt	✗	Disentangled Representation	DST-Attention & Harmonic-RoPE	+2.7% Acc (ST-LLM)
PaMi-VDPO (Ding et al., 2025)	arXiv'25	Scn-Evt	✗	Preference Optimization	Part-mismatch visual negatives	+5.9% Acc (LLaVA-OenVision)
MMA (Dang et al., 2025)	IJCAI'25	Scn-Evt	✗	Parameter-Efficient Tuning	Dual-path visual-text alignment	+2.0% Acc (MA-LMM)
VistaDPO (Huang et al., 2025a)	ICML'25	O-A, S-E	✗	Visual-State DPO	Penalizing low visual-dependency tokens	+36.5% Acc (Video-LLaVA)
VideoHallu-GRPO (Li et al., 2025c)	NeurIPS'25	O-A, S-E	✗	RL Fine-Tuning	Group-relative rewards on counter-intuitive data	+4.7% Acc (Qwen2.5-VL)
Audio-Visual Conflict Mitigation (Content Fabrication)						
AVHModel-Align-FT (Sung-Bin et al., 2025)	ICLR'25	Act	✗	Instruction Tuning	Modality-Disentangled Data	+33.8% Acc (Video-LLaMA)
AVCD (Jung et al., 2025)	NeurIPS'25	Act	✓	Trimodal Contrastive Decoding	Dominance-aware Attentive Masking	+1.6% Acc (VideoLLaMA2)
PEP-MEK (Xing et al., 2025)	arXiv'25	Emo	✓	Predict-Explain-Predict	Knowledge Extraction & Refinement	+9.1% Acc (Gemini2.5-Flash)

phasize motion cues over object presence. SANTA (Chang et al., 2025) applies fine-grained contrastive tuning using hard negatives where actions differ but entities remain the same, encouraging the model to rely on motion rather than co-occurrence patterns. TCD (Zhang et al., 2024) suppresses object-triggered predictions by subtracting logits from temporally shuffled inputs, guiding the model to attend to dynamic cues rather than static priors.

For *scene–event hallucination*, methods reduce the influence of background context on event inference. MASH-VLM (Bae et al., 2025) uses disentangled spatial-temporal attention to prevent the model from predicting actions based solely on static backgrounds. PaMi-VDPO (Ding et al., 2025) introduces preference learning with visually mismatched negatives to train the model to verify event descriptions against the actual scene. MMA (Dang et al., 2025) aligns local visual details with textual tokens through a dual-path adapter, reinforcing grounding in specific cues over general context.

Some methods address both cases. VistaDPO (Huang et al., 2025a) penalizes predictions that lack visual grounding by optimizing against prior-driven outputs, encouraging reliance on directly observable evidence. VideoHallu-GRPO (Li et al., 2025c) improves grounding using synthetic videos with counterintuitive scenarios, optimizing model behavior through group-based relative rewards that favor visually consistent responses over learned priors.

Audio-visual conflict mitigation addresses errors caused by dominant audio signals overriding visual input. This area remains underexplored, with few targeted methods.

For *action attribution*, AVHModel-Align-FT (Sung-Bin et al., 2025) fine-tunes on annotations separating audio and visual events, helping models distinguish between auditory and visual sources. AVCD (Jung et al., 2025) uses contrastive learning with modality masking to suppress misleading cues and improve cross-modal grounding.

For *emotion inference*, PEP-MEK (Xing et al., 2025) enforces modality-specific reasoning by requiring models to explain visual evidence before integrating it with audio, reducing overreliance on vocal tone.

5.3 Discussion: Coverage and Trade-offs

Table 2 reveals uneven progress across hallucination types. While Spatiotemporal Dynamics and Context-Driven Fabrication are well addressed, Referential Inconsistency and Audio-Visual Conflict remain underexplored. A key trade-off exists between effectiveness and deployment cost. Training-based methods (e.g., VistaDPO (Huang et al., 2025a), AVHModel-Align-FT (Sung-Bin

et al., 2025)) offer substantial gains (up to 30–36%) by reshaping model priors via reinforcement learning or instruction tuning, but require high training overhead. In contrast, training-free strategies (e.g., SEASON (Wu et al., 2025a), SmartSight (Sun et al., 2025), TCD (Zhang et al., 2024)) are model-agnostic and easier to adopt, though with more limited gains. Mechanism suitability often aligns with error type: inference-time decoding or attention adjustments help correct temporal and logical inconsistencies (e.g., VideoPLR (Pu et al., 2025)), while hallucinations driven by strong priors call for deeper disentanglement during training (e.g., MASH-VLM (Bae et al., 2025)). Reducing the latency of current inference-time methods is critical for real-time and safety-sensitive deployment.

6 Future Directions

Building on the taxonomy in Section 3, we propose two core directions for enhancing hallucination robustness in Vid-LLMs, targeting the underlying causes of dynamic distortion and content fabrication.

6.1 Addressing Dynamic Distortion: Temporal and Referential Fidelity

Dynamic distortion primarily results from a gap between visual encoding and temporal reasoning. This issue is often rooted in the use of static image encoders and pooling-based connectors, which tend to discard motion cues critical for capturing temporal dynamics (Li et al., 2025b, 2023b; Zhang et al., 2023b). As videos become longer, this problem is further exacerbated by the limited capacity of current models to maintain long-range context, leading to semantic drift and referential inconsistency (Xiao et al., 2024; Guo et al., 2025).

To address these limitations, future architectures should adopt video-oriented designs that preserve temporal structure throughout the pipeline. This includes using video-native encoders such as VideoMAE (Wang et al., 2023) and motion-aware connectors that incorporate signals like optical flow (Liu et al., 2025b) to capture velocity and trajectory. For long-range consistency, models may benefit from structured memory mechanisms, including state space models like Mamba (Li et al., 2024) or episodic memory (Wang et al., 2025), to retain persistent entity information over extended sequences.

6.2 Mitigating Content Fabrication: Grounding and Alignment

Content fabrication arises when pretraining priors dominate over visual grounding. Models may hallucinate actions or events based on static entities or scenes, ignoring temporal evidence (Chang et al., 2025; Bae et al., 2025). The problem is worsened by imbalanced modality integration, where dominant audio cues override visual input, leading to cross-modal conflicts (Sung-Bin et al., 2025; Leng et al., 2024).

To reduce fabrication, models should learn to separate priors from perceptual evidence. This can be achieved by using counterfactual strategies, such as introducing negative samples with implausible object–action pairs and applying debiasing objectives that encourage reliance on motion cues (Qi et al., 2024). In audio-visual settings, models should verify visual input before incorporating audio signals to avoid hallucinations caused by sound (Mo and Song, 2024).

7 Conclusion

Video large language models (Vid-LLMs) have achieved significant progress in video-language modeling, but also give rise to unique hallucination patterns that differ from those in static image tasks. This survey introduces a mechanism-based taxonomy that categorizes hallucinations in Vid-LLMs into two primary categories: Dynamic Distortion, referring to the misinterpretation of spatiotemporal progression or referential consistency; and Content Fabrication, referring to ungrounded outputs influenced by statistical context priors or dominant auditory cues. Although recent studies have advanced benchmarking and mitigation of spatiotemporal and context-driven hallucinations, challenges such as referential inconsistency and audio-visual conflicts remain underexplored. Furthermore, most existing mitigation strategies are applied at inference time or as post-training adjustments, highlighting the need for scalable, training-time alignment methods. To improve the robustness of Vid-LLMs, we advocate future research toward developing video-native encoders that preserve motion cues, integrating explicit memory mechanisms to support long-term temporal grounding, and employing counterfactual learning to disentangle model reasoning from prior-driven associations. These directions will be essential for building trustworthy and temporally faithful video-language systems.

Limitations

Previous surveys on hallucination in MLLMs have extended the scope from LLMs and image-based VLMs to include video and other modalities. However, their coverage of video hallucination remains limited, with only brief mentions of benchmarks and mitigation efforts, lacking structured categorization or causal analysis. This survey addresses this gap by presenting a mechanism-driven taxonomy of hallucinations in Vid-LLMs. We introduce a layered classification framework, review recent studies in greater depth, analyze the underlying causes of hallucinations, and outline future research directions. While we have strived to cover key developments in Vid-LLM hallucination, some relevant work may be omitted. This survey includes research published up to January 2026.

Ethics Statement

This survey adheres to established ethical standards for academic research. All referenced works are publicly available, and no human subjects or personally identifiable information are involved. The purpose of this survey is to facilitate academic understanding and encourage responsible development in the study of hallucination in Vid-LLMs. All prior work has been properly cited, with appropriate credit given to original contributions.

References

Kyungho Bae, Jinhyung Kim, Sihaeng Lee, Soonyoung Lee, Gunhee Lee, and Jinwoo Choi. 2025. MASH-VLM: mitigating action-scene hallucination in video-llms through disentangled spatial-temporal representations. In *CVPR*, pages 13744–13753. Computer Vision Foundation / IEEE.

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

Jianfeng Cai, Wengang Zhou, Zongmeng Zhang, Jiale Hong, Nianji Zhan, and Houqiang Li. 2025. Mitigating hallucination in videollms via temporal-aware activation engineering. *CoRR*, abs/2505.12826.

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

Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. 2024. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *ICRA*, pages 14093–14100.

Wey Yeh Choong, Yangyang Guo, and Mohan S. Kankanhalli. 2024. Vidhal: Benchmarking temporal hallucinations in vision llms. *CoRR*, abs/2411.16771.

Jisheng Dang, Shengjun Deng, Haochen Chang, Teng Wang, Bimei Wang, Shude Wang, Nannan Zhu, Guo Niu, Jingwen Zhao, and Jizhao Liu. 2025. Hallucination reduction in video-language models via hierarchical multimodal consistency. In *IJCAI*, pages 9167–9175. ijcai.org.

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

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, Ran He, Rongrong Ji, Yunsheng Wu, Caifeng Shan, and Xing Sun. 2024. VITA: towards open-source interactive omni multimodal LLM. *CoRR*, abs/2408.05211.

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

Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. 2025. VTG-LLM: integrating timestamp knowledge into video llms for enhanced video temporal grounding. In *AAAI*, pages 3302–3310. AAAI Press.

Haojian Huang, Haodong Chen, Shengqiong Wu, Meng Luo, Jinlan Fu, Xinya Du, Hanwang Zhang, and Hao Fei. 2025a. Vistadpo: Video hierarchical spatial-temporal direct preference optimization for large video models. In *ICML*. OpenReview.net.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55.

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

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. 2025d. Online video understanding: Ovbench and videochat-online. In *CVPR*,

pages 3328–3338. Computer Vision Foundation / IEEE.

Chaeyoung Jung, Youngjoon Jang, and Joon Son Chung. 2025. AVCD: mitigating hallucinations in audio-visual large language models through contrastive decoding. *CoRR*, abs/2505.20862.

Ming Kong, Xianzhou Zeng, Luyuan Chen, Yadong Li, Bo Yan, and Qiang Zhu. 2025. Mhbench: Demystifying motion hallucination in videollms. In *AAAI*, pages 4401–4409. AAAI Press.

Wei Lan, Wenyi Chen, Qingfeng Chen, Shirui Pan, Huiyu Zhou, and Yi Pan. 2024. A survey of hallucination in large visual language models. *CoRR*, abs/2410.15359.

Kyuho Lee, Euntae Kim, Jinwoo Choi, and Buru Chang. 2025. Noah: Benchmarking narrative prior driven hallucination and omission in video large language models. *arXiv preprint arXiv:2511.06475*.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *CoRR*, abs/2410.12787.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. 2025a. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *CVPR*, pages 13723–13733. Computer Vision Foundation / IEEE.

Jinxuan Li, Chaolei Tan, Haoxuan Chen, Jianxin Ma, Jian-Fang Hu, Wei-Shi Zheng, and Jianhuang Lai. 2025b. Image-to-video transfer learning based on image-language foundation models: A comprehensive survey. *CoRR*, abs/2510.10671.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023a. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. 2024. Videomamba: State space model for efficient video understanding. In *ECCV*, volume 15084, pages 237–255. Springer.

Zongxia Li, Xiyang Wu, Guangyao Shi, Yubin Qin, Hongyang Du, Tianyi Zhou, Dinesh Manocha, and Jordan Lee Boyd-Graber. 2025c. Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding. *CoRR*, abs/2505.01481.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, pages 5971–5984. Association for Computational Linguistics.

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

Hui Liu and Xiaojun Wan. 2023. Models see hallucinations: Evaluating the factuality in video captioning. In *EMNLP*, pages 11807–11823. Association for Computational Linguistics.

Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. 2025a. VALOR: vision-audio-language omni-perception pretraining model and dataset. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(2):708–724.

Ruyang Liu, Shangkun Sun, Haoran Tang, Wei Gao, and Ge Li. 2025b. Flow4agent: Long-form video understanding via motion prior from optical flow. In *CVPR*, pages 23817–23827.

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

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

Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. 2024. Vista-llama: Reducing hallucination in video language models via equal distance to visual tokens. In *CVPR*, pages 13151–13160. IEEE.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL (1)*, pages 12585–12602. Association for Computational Linguistics.

Shentong Mo and Yibing Song. 2024. Aligning audio-visual joint representations with an agentic workflow. In *NeurIPS*.

Chirag Parikh, Deepti Rawat, Rakshitha R. T, Tathagata Ghosh, and Ravi Kiran Sarvadevabhatla. 2025. Roadsocial: A diverse videoqa dataset and benchmark for road event understanding from social video narratives. In *CVPR*, pages 19002–19011. Computer Vision Foundation / IEEE.

Bowei Pu, Chuanbin Liu, Yifan Ge, Peichen Zhou, Yiwei Sun, Zhiyin Lu, Jiankang Wang, and Hongtao Xie. 2025. Alternating perception-reasoning for hallucination-resistant video understanding. *CoRR*, abs/2511.18463.

Zhaobo Qi, Yibo Yuan, Xiaowen Ruan, Shuhui Wang, Weigang Zhang, and Qingming Huang. 2024. Bias-conflict sample synthesis and adversarial removal debias strategy for temporal sentence grounding in video. In *AAAI*, pages 4533–4541. AAAI Press.

Ruchit Rawal, Reza Shirkavand, Heng Huang, Gowthami Somepalli, and Tom Goldstein. 2025. ARGUS: hallucination and omission evaluation in video-llms. *CoRR*, abs/2506.07371.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *EMNLP (Findings)*, pages 11709–11724. Association for Computational Linguistics.

Ashish Seth, Utkarsh Tyagi, Ramaneswaran Selvakumar, Nishit Anand, Sonal Kumar, Sreyan Ghosh, Ramani Duraiswami, Chirag Agarwal, and Dinesh Manocha. 2025. EGOILLUSION: Benchmarking hallucinations in egocentric video understanding. In *EMNLP*, pages 28461–28480.

Li Sun, Liuan Wang, Jun Sun, and Takayuki Okatani. 2024. Temporal insight enhancement: Mitigating temporal hallucination in video understanding by multimodal large language models. In *ICPR (7)*, volume 15307 of *Lecture Notes in Computer Science*, pages 455–473. Springer.

Yiming Sun, Mi Zhang, Feifei Li, Geng Hong, and Min Yang. 2025. Smartsight: Mitigating hallucination in video-llms without compromising video understanding via temporal attention collapse. *CoRR*, abs/2512.18671.

Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. 2025. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. In *ICLR*. OpenReview.net.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. Videomae V2: scaling video masked autoencoders with dual masking. In *CVPR*, pages 14549–14560. IEEE.

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. 2024a. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV (85)*, volume 15143 of *Lecture Notes in Computer Science*, pages 396–416. Springer.

Yun Wang, Long Zhang, Jingren Liu, Jiaqi Yan, Zhanjie Zhang, Jiahao Zheng, Xun Yang, Dapeng Wu, Xiangyu Chen, and Xuelong Li. 2025. Episodic memory representation for long-form video understanding. *CoRR*, abs/2508.09486.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. 2024b. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *CoRR*, abs/2406.16338.

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

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

Zhenyu Wu, Ziwei Wang, Xiuwei Xu, Jiwen Lu, and Haibin Yan. 2023. Embodied task planning with large language models. *CoRR*, abs/2307.01848.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *ICLR*. OpenReview.net.

Bohao Xing, Xin Liu, Guoying Zhao, Chengyu Liu, Xiaolan Fu, and Heikki Kälviäinen. 2025. Emotionhallucer: Evaluating emotion hallucinations in multimodal large language models. *CoRR*, abs/2505.11405.

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. 2024. Vript: A video is worth thousands of words. In *NeurIPS*.

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

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134. AAAI Press.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP (Demos)*, pages 543–553. Association for Computational Linguistics.

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP (Demos)*, pages 543–553. Association for Computational Linguistics.

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

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023c. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.

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

Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *AAAI*, pages 7590–7598. AAAI Press.