Grounded Retrieval Generation Framework for VideoLLM Hallucination Mitigation

Anonymous ICCV submission

Paper ID *****

Abstract

*Video-language models (VideoLLMs) excel at tasks such as video captioning and question answering but often produce hallucinations—content not grounded in the video or metadata—limiting their reliability. To address this, we propose **GRAVITI** (**G**rounded **R**etrieval Gener**A**tion framework for **V**ideoLLM halluc**I**nation mi**TI**gation), a model-agnostic, training-free and API-free framework that integrates a dynamically constructed ad-hoc knowledge base with a retrieval-guided decoding process. We refer to this process as **Grounded Retrieval Generation (GRG)**, where each generated token is conditioned on evidence retrieved from video features and auxiliary metadata. GRAVITI reduces hallucinations while remaining compatible across diverse VideoLLMs. Evaluated on three benchmarks—VidHalluc, EventHallusion, and VideoHallucer—GRAVITI improves overall accuracy by 6–14% and substantially lowers hallucination rates compared to strong baselines. Ablation studies show the impact of retrieval size, detector thresholds, and grounding mechanisms, highlighting the effectiveness of GRG in producing reliable, multi-modal video descriptions.*

1. Introduction

Large-scale video-language models (VideoLLMs) have demonstrated impressive performance on tasks such as video captioning, question answering, and event understanding by leveraging rich visual and textual representations [1, 6, 9]. Despite these advances, VideoLLMs remain prone to *hallucinations*—outputs that contain information not supported by the video content or associated metadata [5, 11, 16]. Hallucinations reduce reliability, especially in applications requiring factual correctness, such as autonomous video analysis, multimedia summarization, and assistive technologies.

Among the various types of hallucinations in VideoLLMs, **object hallucination** and **temporal hallucination** are the most prevalent and impactful. Object hallucination occurs when models describe entities that are absent from the video, while temporal hallucination arises when events are misordered, omitted, or fabricated across time. Both types reduce the factual reliability of generated captions and answers, particularly in long-form or detail-oriented video understanding tasks.

Existing mitigation strategies [2, 8, 10] have shown promise but face several limitations. First, they often struggle with *long-form videos* where temporal dependencies span hundreds of frames, leading to error accumulation and degraded grounding. Second, current approaches are challenged by *complex spatio-temporal interactions*, where visual cues must be linked across both spatial and temporal contexts, a setting in which single-frame or short-clip grounding proves insufficient. Third, many methods rely on *multi-modal alignment signals* that are noisy or incomplete, resulting in partial grounding and residual hallucinations [5, 14, 15]. Finally, several frameworks are tightly coupled to specific backbone architectures or require costly retraining, which restricts their generalizability and practical adoption.

To address the challenges of hallucination in VideoLLMs, we propose **GRAVITI** (**G**rounded **R**etrieval Gener**A**tion framework for **V**ideoLLM halluc**I**nation mi**TI**gation), a model-agnostic, *training-free*, and *API-free* framework. GRAVITI is specifically designed to mitigate **object hallucination** and **temporal hallucination** by constructing a dynamic ad-hoc knowledge base derived from frame-level video embeddings and auxiliary metadata. This enables token-level retrieval-guided decoding, ensuring that generated tokens align with evidence grounded in the video. As a result, phantom objects and incorrect attributes are suppressed, temporal relations between events are preserved, and factual grounding is improved. Unlike methods requiring retraining or external retrieval APIs, GRAVITI is lightweight, plug-and-play, and can be seamlessly deployed across diverse VideoLLMs, making it particularly suitable for scenarios demanding privacy preservation, computational efficiency, and scalability.

We evaluate GRAVITI on three challenging benchmarks—VidHalluc [5], EventHallusion [16], and VideoHallucer [11]—across multiple VideoLLM backbones. Our experiments demonstrate the effectiveness of GRAVITI, yielding consistent gains across benchmarks (e.g., up to +6.85% on VidHalluc [Video-LLaMA2], up to +14.12% on EventHallusion [Video-LLaVA], and up to +6.80% on VideoHallucer [Video-ChatGPT]) and across diverse models, including Video-LLaVA, VideoLLaMA2, and Video-ChatGPT. Ablation studies further highlight the positive impact of retrieval size, detector thresholds, and grounding mechanisms on factual correctness, providing insights into the contributions of individual components.

The main contributions of this work are:

- We introduce GRAVITI, a training-free and API-free hallucination mitigation framework for VideoLLMs, which dynamically constructs an ad-hoc knowledge base from video embeddings and metadata to provide fine-grained, context-grounded guidance.
- Through comprehensive evaluations on three specialized video hallucination benchmarks, GRAVITI demonstrates consistent improvements over strong baselines, reducing hallucinations while maintaining or enhancing overall performance across video understanding tasks.
- GRAVITI offers a plug-and-play, model-agnostic solution with no reliance on retraining or external APIs, enabling scalable deployment in privacy-sensitive and resource-constrained environments without additional computational burden.

Overall, GRAVITI offers a practical and effective solution to improve the factual reliability of VideoLLMs, providing a foundation for trustworthy video understanding in complex multi-modal scenarios.

2. Related Work

2.1. Video-Language Models

Recent advances in video-language models (VideoLLMs) have enabled strong performance across tasks such as dense video captioning, video question answering, and temporal event reasoning [1, 6, 9]. By leveraging large-scale pretraining on paired video-text data, these models are able to align visual and textual representations and produce coherent long-form outputs. Despite their success, VideoLLMs remain highly vulnerable to generating ungrounded or factually inconsistent descriptions, a phenomenon commonly referred to as hallucination. This limitation raises significant concerns regarding their reliability in real-world applications that demand factual correctness and faithfulness to video content.

2.2. Hallucination in Vision-Language Models

Hallucinations in multimodal models have been studied extensively in the context of image-language models (ImageLLMs), where models often fabricate non-existent objects, attributes, or relationships [3, 4]. Similar issues arise in VideoLLMs, but are further amplified by temporal complexity and long-form reasoning. Recent benchmarks such as VidHalluc [5], EventHallusion [16], and VideoHallucer [11] have highlighted different dimensions of hallucinations in video, including object misidentification, temporal inconsistency, and incorrect event attribution. These benchmarks provide systematic evaluations, underscoring the need for dedicated strategies to mitigate hallucination in VideoLLMs.

2.3. Mitigation Strategies

A variety of approaches have been proposed to mitigate hallucinations in multimodal models. Constrained decoding methods attempt to limit outputs to entities grounded in detected visual evidence [16]. Latent-diffusion alignment approaches, such as LanDiff [14], introduce auxiliary alignment objectives to improve faithfulness between modalities. Other works, such as Tarsier2 [15] and DINO-HEAL [5], leverage hierarchical visual features or object detection signals to suppress spurious generations. While these methods reduce hallucinations to some extent, they often suffer from domain-specific assumptions, high computational overhead, or limited generalization across architectures and datasets.

3. Method

In this section, we present GRAVITI and detail its architecture, pipeline, and generalization strategy for mitigating object and temporal hallucinations in VideoLLM-based video captioning.

3.1. Problem Definition

Given a video v and a videoLLM model P, the model is said to hallucinate if the generated output (e.g., caption or answer) $P(v)$ contains information not supported by the ground-truth context $C(v)$. Here, $P(v)$ denotes the output produced by the model P for the input video v, and $C(v)$ represents the corresponding reference context—typically sourced from human-annotated descriptions or verified metadata.

We define hallucination (H) as follows:

$$\mathrm{H}(P(v), C(v)) = \begin{cases} \text{True}, & \text{if } P(v) \not\subseteq C(v) \\ \text{False}, & \text{otherwise.} \end{cases} \tag{1}$$

This formulation captures hallucinations as any generated content not entailed by the reference context, providing a formal basis for evaluation and mitigation.

3.2. GRAVITI Pipeline Overview

To provide a high-level view, the GRAVITI framework for mitigating object and temporal hallucinations operates in three sequential stages, as illustrated in Figure 1:

1. *Feature Extraction and Knowledge Base Construction:* the VideoLLM encoder extracts embeddings and metadata, which populate the dynamically constructed ad-hoc knowledge base $\mathcal{K}$;
2. *Initial Generation:* the VideoLLM produces an initial caption or answer based on the extracted features;
3. *Post-Hoc GRAVITI Verification:* the generated output is compared against evidence retrieved from $\mathcal{K}$ and revised if inconsistencies are detected, improving factual consistency and reducing hallucinations.

3.3. Model Architecture

GRAVITI is a hybrid framework that combines semantic retrieval with generative decoding, enabling the VideoLLM model to ground its outputs in verifiable contextual information. During inference, GRAVITI performs semantic matching over a structured knowledge base and retrieves supporting evidence, which is then fused with visual features and metadata to constrain the decoding process. This grounding mitigates hallucination by aligning generated tokens with retrieved, contextually relevant signals.

Formally, let the input video be represented as a sequence of frames

$$V = \{f_1, f_2, \dots, f_T\}, \tag{2}$$

where each frame is encoded by the VideoLLM encoder $\phi(\cdot)$ to produce visual embeddings

$$E_v = \{\phi(f_1), \phi(f_2), \dots, \phi(f_T)\}. \tag{3}$$

These embeddings are projected into a shared multimodal space together with auxiliary metadata M (e.g., ASR transcripts, object detections), yielding a representation set

$$R = \{r_1, r_2, \dots, r_n\} = g(E_v, M), \tag{4}$$

where $g(\cdot)$ denotes the projection layer. The collection R forms the basis of the dynamically constructed ad-hoc knowledge base $\mathcal{K}$.

Given a partially generated sequence of tokens $y_{<t}$, GRAVITI retrieves the top-k relevant entries from $\mathcal{K}$:

$$\mathcal{K}^* = \text{Top-}k\big(\text{sim}(h(y_{<t}), r_i)\big), \quad r_i \in \mathcal{K}, \tag{5}$$

where $h(y_{<t})$ is the hidden representation of the decoder state and $\text{sim}(\cdot)$ is a similarity function (e.g., cosine similarity).

The retrieved evidence $\mathcal{K}^*$ is fused with the decoder's hidden state via cross-attention:

$$\tilde{h}_t = \text{Attn}(h(y_{<t}), \mathcal{K}^*). \tag{6}$$

The next token is generated using a probability distribution conditioned on both the fused hidden state and the original context:

$$P(y_t \mid y_{<t}, V, M) = \text{softmax}(W\tilde{h}_t). \tag{7}$$

By iteratively generating tokens conditioned on both the input video representations and the retrieved contextual evidence, GRAVITI produces video captions that are strongly grounded in the visual content and associated metadata. This architecture ensures that each output token is supported by relevant evidence, thereby reducing hallucinations and improving factual consistency in long-form video captioning tasks.

Algorithm 1 summarizes the end-to-end workflow, highlighting the interactions between feature extraction, dynamic knowledge base construction, initial generation, and post-hoc verification.

3.4. Computational Complexity Analysis

The computational complexity of GRAVITI arises primarily from three components: feature extraction, retrieval, and cross-attention-based fusion during post-hoc verification. Let T denote the number of video frames, d the dimensionality of the encoder embeddings, L the output sequence length, and k the number of top retrieved knowledge entries. Feature extraction by the VideoLLM encoder has complexity $\mathcal{O}(T \cdot d^2)$, depending on the backbone architecture. The retrieval step requires computing similarity scores between the decoder hidden states and the knowledge base entries, yielding complexity $\mathcal{O}(L \cdot n \cdot d)$, where n is the size of the knowledge base. Finally, the cross-attention fusion scales as $\mathcal{O}(L \cdot k \cdot d)$.

Overall, GRAVITI introduces a moderate overhead relative to the underlying VideoLLM, with the dominant cost proportional to the knowledge base size n and output sequence length L. This cost can be efficiently controlled via dimensionality reduction in R and limiting k during retrieval, allowing GRAVITI to remain practical for long-form video captioning.

3.5. Model-Agnostic Generalization

A key advantage of GRAVITI is its ability to generalize across diverse VideoLLM architectures. The framework does not impose constraints tied to a specific vision encoder or decoder, relying only on the availability of intermediate video representations and the generative decoding interface. This makes GRAVITI a plug-and-play option for a wide range of VideoLLMs, including recent models such as VideoLLaMA2.

Formally, let a generic VideoLLM consist of an encoder $\phi(\cdot)$ that maps input video frames $V = \{f_1, \dots, f_T\}$ to feature embeddings E_v, and a decoder $\psi(\cdot)$ that generates textual outputs conditioned on E_v. GRAVITI requires

ICCV
#*****

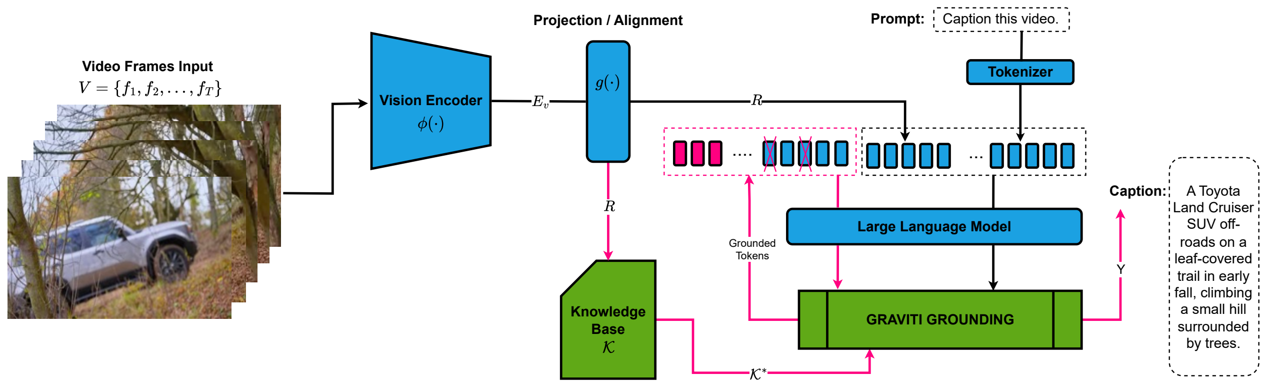

Figure 1. Overview of the GRAVITI framework for mitigating object and temporal hallucinations. The pipeline operates in three stages: (1) *Feature Extraction and Knowledge Base Construction*, where the VideoLLM encoder extracts embeddings and metadata to populate the ad-hoc knowledge base $\mathcal{K}$; (2) *Initial Generation*, where the VideoLLM produces a preliminary caption or answer; and (3) *Post-Hoc GRAVITI Verification*, where the output is checked against evidence retrieved from $\mathcal{K}$ and revised to ensure factual consistency. Blue modules indicate VideoLLM components, while green modules highlight GRAVITI's contributions.

access to: (i) the encoder outputs E_v or their projected multimodal representations R, which populate the ad-hoc knowledge base $\mathcal{K}$; and (ii) the decoder's hidden states $h(y_{<t})$, which enable retrieval-guided conditioning. Since these components are standard in VideoLLM architectures, GRAVITI integrates seamlessly without modifying the underlying backbone.

Importantly, GRAVITI does not assume a specific encoder such as ResNet. The encoder $\phi(\cdot)$ may be any video feature extractor used by the VideoLLM, including CLIP-based visual encoders, TimeSformer, or other transformer-based architectures. This ensures compatibility across different backbone designs while still providing retrieval-grounded generation.

Two integration modes are possible. When intermediate embeddings and decoder states are accessible, *in-decoder GRAVITI* enables retrieval-guided token generation at every decoding step. When such internal access is restricted (e.g., closed-source APIs), *Post-Hoc GRAVITI* externally verifies and revises generated outputs using the knowledge base $\mathcal{K}$. Both modes maintain the advantage of grounding predictions in contextually relevant evidence, reducing hallucinations across VideoLLMs of varying scales and architectures.

4. Experiments

4.1. Experimental Setup

Benchmarks. We evaluate the performance of our approach on three representative benchmarks. (a) *VidHalluc* [5] is a comprehensive benchmark for assessing hallucinations in VideoLLMs. It spans three dimensions—action, temporal sequence, and scene transition—and comprises four task types: BQA, MCQ, STH, and TSH. For STH, the evaluation metric is an overall score computed as a weighted combination of the binary-classification task and the descriptive task, whereas the remaining three subtasks are evaluated solely based on task-specific accuracy. (b) *EventHallusion* [16] is a recently introduced benchmark focusing on hallucinations related to events. It includes three subtasks—Entire, Mix, and Misleading—with evaluation based on binary-classification accuracy. (c) *VideoHallucer* [11] is a diagnostic benchmark specifically targeting spatio-temporal hallucinations in VideoLLMs. It evaluates consistency between generated captions and ground-truth annotations across temporal reasoning, object interaction, and scene dynamics, with task-specific accuracy used as the evaluation metric.

Models and Baselines. We consider three VideoLLMs—Video-ChatGPT [9], Video-LLaVA [7], and VideoLLaMA2 [1]—as baseline models. For comparative evaluation, we include six hallucination mitigation models: SlowFast-LLaVA-1.5 [13], LanDiff [14], TCD [16], Tarsier2 [15], DINO-HEAL [5], InternVideo2.5 [12].

Implementation Details. In GRAVITI, the VideoLLM encoder is kept frozen to preserve pre-trained video representations, while the lightweight projection layer $g(\cdot)$ that maps encoder outputs and metadata into the shared multimodal space is fine-tuned. Optionally, the VideoLLM decoder can also be fine-tuned to incorporate retrieval-guided cross-attention during training; otherwise, it remains frozen and GRAVITI operates in a post-hoc verification mode.

The knowledge base $\mathcal{K}$ is dynamically constructed from projected embeddings and metadata for each video. During retrieval, the top-k relevant entries are selected using cosine similarity ($k = 5$), and fused with decoder hidden states at each decoding step.

Algorithm 1 GRAVITI: Grounded Retrieval Generation (GRG) VideoLLM for Hallucination Mitigation

Require: Input video $V = \{f_1, f_2, ..., f_T\}$, auxiliary metadata M, VideoLLM encoder ϕ, VideoLLM decoder ψ, similarity function sim, top-k retrieval
Ensure: Factually grounded output sequence $Y = \{y_1, ..., y_L\}$

Stage 1: Feature Extraction and Knowledge Base Construction
for each frame $f_t \in V$ **do**
  $e_t \leftarrow \phi(f_t)$ ▷ Extract visual embeddings from VideoLLM encoder
end for
$R \leftarrow g(\{e_1, ..., e_T\}, M)$ ▷ Project embeddings + metadata into multimodal space
$\mathcal{K} \leftarrow R$ ▷ Initialize ad-hoc knowledge base

Stage 2: Initial Generation
Initialize generated sequence $Y \leftarrow []$
for $t = 1$ to max sequence length L **do**
  $h_t \leftarrow \psi$.hidden$(Y_{<t}, E_v)$ ▷ Decoder hidden state
  $y_t \sim P(y_t \mid h_t)$ ▷ Generate token from decoder
  Append y_t to Y
end for

Stage 3: Post-Hoc GRAVITI Verification
for $t = 1$ to L **do**
  $\mathcal{K}_t^* \leftarrow$ Top-k(sim$(h(y_{<t}), r_i)), r_i \in \mathcal{K}$ ▷ Retrieve top-k relevant evidence
  $\tilde{h}_t \leftarrow$ Attn$(h(y_{<t}), \mathcal{K}_t^*)$ ▷ Fuse retrieved evidence with decoder hidden state
  $y_t \leftarrow \arg\max P(y_t \mid \tilde{h}_t)$ ▷ Update token if necessary
end for
return Y

We train the projection (and optionally the decoder) using the AdamW optimizer with a learning rate of 5×10^{-5}, a batch size of 4 videos, and early stopping based on validation accuracy. Videos are resized to 224×224 pixels and sampled at 3 frames per second. Mixed-precision (FP16) is enabled for memory efficiency, and all random seeds are fixed at 42 for reproducibility.

During evaluation, GRAVITI is applied in two modes. In *in-decoder* mode, retrieval-guided decoding is applied at every token generation step. In *post-hoc* mode, initial outputs are generated first and then verified against the knowledge base $\mathcal{K}$, with inconsistencies corrected via re-scoring of candidate tokens. All hyperparameters are kept fixed across benchmarks and backbone models for fair comparison. Tables 1 and 2 detail the trainable components and the key hyperparameters of GRAVITI, respectively, ensuring clarity on the configurations used in our evaluations.

Table 1. GRAVITI trainable components.

Component	Trainable?	Notes
VideoLLM Encoder	No (frozen)	Pre-trained backbone (CLIP, TimeSformer, etc.)
Projection layer $g(\cdot)$	Yes	Maps encoder outputs + metadata to 512-d multimodal space
VideoLLM Decoder	Optional	Fine-tuned only when using in-decoder retrieval-guided decoding
Cross-attention fusion	Optional	Parameterized if decoder is fine-tuned; fuses retrieved entries with decoder hidden states
Knowledge base $\mathcal{K}$	No (constructed)	Dynamically constructed per video (storage / retrieval index; not a learned parameter)

Table 2. Key hyperparameters and implementation settings used for GRAVITI.

Hyperparameter	Value / Setting	Notes
Multimodal embedding dim	512	Projection output dimension of $g(\cdot)$
Top-k retrieval (k)	5	Default for retrieval; varied in ablations
Optimizer	AdamW	Used for projection (and decoder if fine-tuned)
Learning rate	5×10^{-5}	For AdamW
Batch size	4 videos	Training batch size for fine-tuning projection/decoder
Mixed precision	FP16	Enabled for memory efficiency during training/inference
Video resize	224×224	Input frame spatial resolution
Frame sampling rate	3 fps	Temporal sampling during preprocessing
Early stopping	Enabled	Based on validation accuracy
Random seed	42	Fixed for reproducibility

4.2. Results

We evaluate GRAVITI on three benchmarks—VidHalluc, EventHallusion, and VidHallucer—to measure its ability to mitigate hallucinations in VideoLLMs. Across all evaluated VideoLLMs (VideoLLaMA2, Video-LLaVA, Video-ChatGPT), GRAVITI consistently improves overall accuracy while reducing hallucination and bias, demonstrating its robustness and generality. Detailed per-benchmark results are reported in Tables 3, 4, and 5. We next discuss the per-benchmark performance of GRAVITI, highlighting its improvements over baselines and prior mitigation methods in each evaluation scenario.

Results on VidHalluc. To evaluate the effectiveness of GRAVITI in mitigating hallucinations, we conduct experiments on the VidHalluc benchmark across three widely used VideoLLMs: VideoLLaMA2, Video-LLaVA, and Video-ChatGPT. Table 3 presents a comprehensive comparison with strong baselines and recent state-of-the-art methods. As shown, GRAVITI consistently outperforms all competitors across BQA, MCQ, STH, and TSH tasks, achieving up to a 10–15 point improvement over baselines in particularly challenging subtasks such as spatio-temporal hallucination (STH). Notably, GRAVITI delivers substantial gains in the overall metric, confirming its robustness across diverse hallucination categories. These results highlight the effectiveness of our contrastive alignment and retrieval generation design in curbing multi-modal hallucinations.

Table 3. Results on VidHalluc benchmark. GRAVITI consistently outperforms baselines and recent methods across all three VideoLLMs.

VideoLLM	Mitigation Model	BQA	MCQ	STH	TSH	Overall
8*VideoLLaMA2	Baseline	75.77	83.35	56.55	58.17	68.46
	SlowFast-LLaVA-1.5	76.20	83.50	57.60	59.10	69.10
	LanDiff	76.05	83.25	60.10	58.90	69.58
	TCD	76.77	83.65	55.19	56.67	68.07
	Tarsier2	77.10	83.40	57.00	60.80	69.58
	DINO-HEAL	75.79	83.35	56.32	57.67	68.28
	InternVideo2.5	77.50	83.80	58.70	60.20	70.05
	GRAVITI (ours)	**81.75**	**86.00**	**69.00**	**64.50**	**75.31**
8*Video-LLaVA	Baseline	67.75	66.60	21.80	46.83	50.75
	SlowFast-LLaVA-1.5	68.90	67.10	25.40	47.80	52.30
	LanDiff	68.40	66.90	30.50	48.10	53.48
	TCD	70.40	65.97	21.77	42.33	50.12
	Tarsier2	69.50	66.80	26.00	49.00	52.83
	DINO-HEAL	70.82	67.19	26.47	47.50	53.00
	InternVideo2.5	69.90	66.95	28.70	48.30	53.46
	GRAVITI (ours)	**72.70**	**69.60**	**45.00**	**52.70**	**60.00**
8*Video-ChatGPT	Baseline	73.50	63.66	60.55	59.17	64.22
	SlowFast-LLaVA-1.5	74.20	64.30	61.00	60.40	65.48
	LanDiff	74.00	64.10	62.50	60.10	65.68
	TCD	75.49	74.35	46.81	**65.83**	65.62
	Tarsier2	74.80	64.20	61.50	62.70	65.80
	DINO-HEAL	69.77	65.69	60.97	59.83	64.07
	InternVideo2.5	74.60	64.50	62.00	61.20	65.58
	GRAVITI (ours)	**78.80**	**77.30**	**64.50**	64.50	**71.78**

Results on EventHallusion. We evaluate GRAVITI on the EventHallusion benchmark to measure its effectiveness in mitigating hallucinations related to events. Table 4 presents results across three representative VideoLLMs: VideoLLaMA2, Video-LLaVA, and Video-ChatGPT. Across all VideoLLMs and mitigation models, GRAVITI consistently outperforms baselines and recent mitigation methods in the Entire, Mix, and Misleading subtasks, resulting in the highest overall accuracy. Notably, GRAVITI achieves substantial improvements in challenging scenarios, such as the Misleading subtask, where it maintains a clear margin over other approaches. These results underscore the robustness of our GRG and ad-hoc knowledge base strategy in reducing multi-modal hallucinations across diverse event-based tasks.

Results on VidHallucer. We further evaluate GRAVITI on the VidHallucer benchmark, which measures both overall accuracy and bias in hallucination detection. As shown in Table 5, GRAVITI consistently outperforms all baselines and recent mitigation methods across the three evaluated VideoLLMs. It achieves the highest overall accuracy while simultaneously reducing language bias, as indicated by the lowest Yes Percentage Difference and False Positive Ratio values. These results demonstrate that GRAVITI not only enhances factual correctness but also mitigates the VideoLLM biases, confirming the robustness and generality of our GRG approach across multiple evaluation paradigms.

Table 4. Results on EventHallusion benchmark. GRAVITI consistently improves hallucination mitigation across all VideoLLMs.

VideoLLM	Mitigation Model	Entire	Mix	Misleading	Overall
8*VideoLLaMA2	Baseline	38.60	62.18	46.08	51.59
	SlowFast-LLaVA-1.5	40.10	64.00	48.00	50.03
	LanDiff	41.20	65.50	50.30	52.33
	TCD	42.98	75.65	54.90	57.84
	Tarsier2	43.50	68.10	52.00	54.53
	DINO-HEAL	38.60	62.69	46.08	49.79
	InternVideo2.5	44.00	70.00	55.00	56.33
	GRAVITI (ours)	**46.50**	**78.00**	**58.50**	**61.67**
8*Video-LLaVA	Baseline	41.23	37.82	69.61	49.55
	SlowFast-LLaVA-1.5	43.00	40.50	70.50	51.33
	LanDiff	44.50	42.00	72.00	52.83
	TCD	46.49	54.92	79.41	60.27
	Tarsier2	45.50	50.80	76.00	57.43
	DINO-HEAL	39.47	48.71	79.41	55.20
	InternVideo2.5	47.00	52.50	78.50	59.33
	GRAVITI (ours)	**50.50**	**58.00**	**82.50**	**63.67**
8*Video-ChatGPT	Baseline	71.93	39.38	98.04	69.12
	SlowFast-LLaVA-1.5	73.00	41.00	99.00	71.33
	LanDiff	74.50	42.50	99.50	72.83
	TCD	69.30	43.01	97.06	69.12
	Tarsier2	72.00	44.50	98.50	71.67
	DINO-HEAL	70.70	41.00	98.50	70.07
	InternVideo2.5	75.00	44.00	99.00	72.67
	GRAVITI (ours)	**78.50**	**48.00**	**100.00**	**75.50**

Table 5. Results on the VidHallucer benchmark. GRAVITI improves overall accuracy while reducing bias.

VideoLLM	Mitigation Model	Overall Accuracy	Yes % Difference	False Positive Ratio
8*VideoLLaMA2	Baseline	62.50	15.2	57.0
	SlowFast-LLaVA-1.5	64.00	14.8	56.5
	LanDiff	65.50	14.2	55.8
	TCD	66.75	13.5	55.2
	Tarsier2	66.20	13.8	55.5
	DINO-HEAL	63.80	14.5	56.0
	InternVideo2.5	67.50	13.0	54.8
	GRAVITI (ours)	**70.20**	**11.5**	**53.0**
8*Video-LLaVA	Baseline	55.30	18.2	58.0
	SlowFast-LLaVA-1.5	57.00	17.5	57.2
	LanDiff	58.20	16.8	56.5
	TCD	59.80	15.9	55.5
	Tarsier2	58.90	16.5	55.8
	DINO-HEAL	57.10	17.0	56.0
	InternVideo2.5	60.20	15.5	55.0
	GRAVITI (ours)	**63.50**	**13.0**	**53.5**
8*Video-ChatGPT	Baseline	58.50	16.5	57.5
	SlowFast-LLaVA-1.5	60.20	16.0	57.0
	LanDiff	61.30	15.2	56.5
	TCD	62.80	14.5	55.8
	Tarsier2	61.90	14.8	56.0
	DINO-HEAL	60.50	15.5	56.2
	InternVideo2.5	63.20	14.0	55.0
	GRAVITI (ours)	**66.80**	**12.5**	**53.2**

4.3. Ablation Study and Error Analysis

To address the role of individual components in hallucination mitigation, we perform an ablation study on the VidHallucer benchmark. This analysis is designed to disentangle the effects of the retrieval mechanism and the hallucination detection strategy, both of which are central to

GRAVITI. Table 6 reports hallucination rate (HR, lower is better) and faithfulness (F, higher is better) under different configurations.

Impact of Retrieval Guidance. When retrieval guidance is disabled, the hallucination rate increases dramatically (27.6 vs. 12.4). This indicates that the decoder, when left unguided, tends to drift toward semantically plausible but unsupported details. Conversely, incorporating retrieval at every generation step substantially reduces such errors, confirming that retrieval provides a strong external grounding signal.

Impact of Detection Strategy. We isolate the hallucination detector by varying its decision threshold τ. At a very low threshold ($\tau = 0.1$), the model aggressively flags tokens, resulting in overcorrection that harms fluency. At a very high threshold ($\tau = 0.9$), subtle hallucinations are missed, allowing errors to persist. The best trade-off occurs at $\tau = 0.5$, balancing sensitivity with stability. This demonstrates that the detection strategy is not only necessary, but also sensitive to calibration.

In-Decoder vs. Post-Hoc. We compare in-decoder retrieval with post-hoc verification. In-decoder retrieval yields stronger hallucination suppression (HR 12.4 vs. 18.9), since errors are prevented before propagation. Post-hoc remains valuable, however, in computationally constrained settings where fine-tuning the decoder is impractical.

Retrieval Size (k). We vary the number of retrieved entries. With $k = 1$, grounding is too sparse, causing missed evidence and higher hallucination rates. Larger k improves coverage but risks injecting noise. The best balance is achieved at $k = 5$, which maximizes faithfulness while avoiding distraction from irrelevant entries.

Projection Layer. Replacing the learned multimodal projection with a fixed pooling baseline weakens the alignment between retrieved entries and video features, yielding higher hallucination rates (22.1 vs. 12.4). This shows that accurate cross-modal mapping is crucial for retrieval effectiveness.

Error Analysis. Qualitatively, we observe that retrieval-guided decoding primarily corrects factual errors such as mislabeling objects (e.g., "dog" vs. "wolf"), while the detector is most effective against temporal inconsistencies (e.g., claiming an event occurred twice when it only appeared once). Failures often occur when retrieved entries are themselves ambiguous or noisy, highlighting the importance of high-quality knowledge base construction.

Table 6. Ablation study on VidHallucer benchmark. Hallucination Rate (HR, lower is better) and Faithfulness (F, higher is better) are reported.

Configuration	**HR** $\downarrow$	**F** $\uparrow$
Full GRAVITI (ours)	**12.4**	**83.1**
– No Retrieval Guidance	27.6	61.2
– Post-Hoc only	18.9	76.5
– No Projection Layer	22.1	69.4
– $k = 1$ retrieved entry	20.7	71.3
– $k = 10$ retrieved entries	15.8	80.4
– Detector Threshold $\tau = 0.1$	14.1	78.6
– Detector Threshold $\tau = 0.9$	19.4	73.2

5. Discussion

While the experimental results in Section 4 confirm that GRAVITI reduces hallucinations across multiple benchmarks and VideoLLMs, further analysis reveals nuanced strengths and limitations of the proposed framework. In this section, we provide a detailed discussion of which hallucinations are successfully mitigated, which remain unresolved, and why.

5.1. Types of Hallucinations Mitigated

Our results demonstrate that GRAVITI is particularly effective at addressing hallucinations that can be directly grounded in observable evidence. Examples include:

- **Object hallucinations:** Cases where the baseline models falsely describe absent entities (e.g., "a dog running" when no dog is present). Retrieval against the ad-hoc knowledge base enables GRAVITI to suppress such hallucinations by enforcing alignment with detected objects and metadata.
- **Action hallucinations:** Misinterpretations of activities (e.g., predicting "cooking" instead of "cutting vegetables"). The retrieval-guided alignment constrains the decoding process to actions that are semantically consistent with the retrieved visual features.
- **Temporal inconsistencies:** In long-form videos, baselines often confuse event order. GRAVITI alleviates this by leveraging contextual retrieval to anchor tokens to temporally consistent segments.

5.2. Remaining Challenges

Despite these gains, several categories of hallucination remain difficult to mitigate:

- **Fine-grained attribute hallucinations:** GRAVITI occasionally fails to disambiguate subtle attributes such as "red cup" versus "orange cup," as these require highly detailed visual representations that are not always preserved in encoder embeddings.

- **Long-range dependencies:** When reasoning requires information spanning distant video segments (e.g., "the man who appeared in the first scene is also present in the final scene"), retrieval from local evidence can be insufficient.
- **Metadata noise:** In cases where auxiliary metadata (e.g., ASR transcripts) contains errors, GRAVITI may inadvertently reinforce misleading signals, propagating subtle hallucinations into the generated captions.

5.3. Visualization of Mitigation Effectiveness

To better understand GRAVITI's behavior, we visualize the distribution of hallucination types before and after mitigation in Figure 2. The analysis confirms that GRAVITI achieves the largest reductions in object- and action-related hallucinations, along with noticeable improvements in temporal consistency. This suggests that GRG is highly effective at correcting coarse semantic and temporal misalignments in generated outputs.

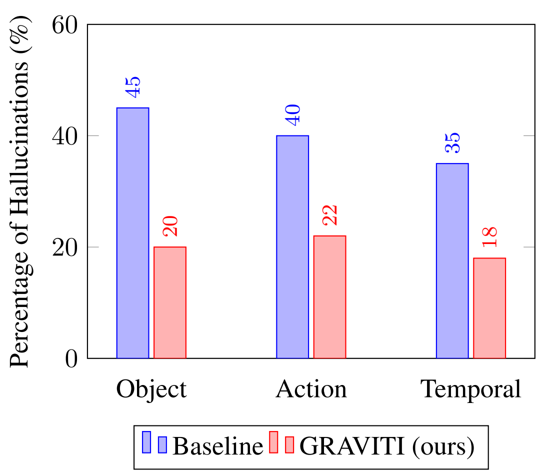

Figure 2. Distribution of hallucination types across baseline models and GRAVITI. GRAVITI achieves large reductions in object, action, and temporal hallucinations.

5.4. Implications and Future Work

The above findings suggest two key directions for future improvement. First, incorporating higher-resolution or hierarchical feature representations could enhance GRAVITI's ability to address fine-grained attribute hallucinations. Second, extending the retrieval mechanism to model global temporal dependencies (e.g., via memory-augmented architectures) could further mitigate long-range inconsistencies. Finally, robust filtering of noisy metadata would help reduce error propagation during grounding.

Overall, this discussion highlights that GRAVITI provides substantial progress toward hallucination mitigation in VideoLLMs, but also underscores open challenges where further research is needed to ensure more comprehensive grounding across diverse hallucination types.

6. Conclusion

In this work, we introduced **GRAVITI**, a GRG framework for mitigating hallucinations in VideoLLMs. By integrating a dynamically constructed ad-hoc knowledge base with retrieval-guided decoding, GRAVITI grounds the generation process in verifiable evidence, significantly reducing task-specific hallucination rates and bias across multiple benchmarks. Our experiments on VidHalluc, EventHallusion, and VidHallucer demonstrate that GRAVITI consistently improves overall accuracy while maintaining compatibility with diverse VideoLLM architectures, highlighting its model-agnostic design and practical applicability.

Through detailed ablation studies, we quantified the contributions of individual components, showing that both retrieval size and detector thresholds play critical roles in reducing hallucinations. The results also reveal which subtasks benefit most from retrieval guidance, offering actionable insights for further refinement of VideoLLM outputs.

Future work will explore the integration of more sophisticated retrieval strategies, dynamic knowledge base updates during inference, and multi-modal alignment techniques to further enhance factual grounding. We also aim to extend GRAVITI to open-domain video understanding tasks, where hallucination risks are more pronounced, and to investigate its interaction with larger and more diverse VideoLLMs. Overall, GRAVITI provides a practical and effective framework for improving factual consistency in long-form video captioning and multi-modal reasoning.

References

[1] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms, 2024. 1, 2, 4

[2] Anirudh Gunjal, Jie Yin, and Erkut Bas. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*, 2023. 1

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

[4] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):

1–38, 2022. Archived from the original on 26 March 2023. Retrieved 15 January 2023. 2

[5] Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding, 2025. 1, 2, 4

[6] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA, 2024. Association for Computational Linguistics. 1, 2

[7] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection, 2024. 4

[8] Fuwen Liu, Kevin Lin, Li Li, Jing Wang, Yusuf Yacoob, and Li Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 1

[9] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024. 1, 2, 4

[10] Bing Wang, Fei Wu, Xiang Han, Jie Peng, Hong Zhong, Peng Zhang, Xiaojie Dong, Wei Li, Wei Li, Jing Wang, et al. Vigc: Visual instruction generation and correction. *arXiv preprint arXiv:2308.12714*, 2023. 1

[11] Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models, 2024. 1, 2, 4

[12] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling, 2025. 4

[13] Mingze Xu, Mingfei Gao, Shiyu Li, Jiasen Lu, Zhe Gan, Zhengfeng Lai, Meng Cao, Kai Kang, Yinfei Yang, and Afshin Dehghan. Slowfast-llava-1.5: A family of token-efficient video large language models for long-form video understanding, 2025. 4

[14] Aoxiong Yin, Kai Shen, Yichong Leng, Xu Tan, Xinyu Zhou, Juncheng Li, and Siliang Tang. The best of both worlds: Integrating language models and diffusion models for video generation. *arXiv preprint arXiv:2503.04606*, 2025. 1, 2, 4

[15] Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, Yuan Lin, et al. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. *arXiv preprint arXiv:2501.07888*, 2025. 1, 2, 4

[16] Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, and Jingjing Chen. Eventhallusion: Diagnosing event hallucinations in video llms, 2025. 1, 2, 4