International Journal of Computer Applications (0975 - 8887)
Volume 187 - No.95, April 2026

TempHalluc-Bench: Evaluating Temporal Hallucination in VideoLLM-based Video Search and Information Extraction

Ahmad Khalil
School of Computer Science,
University of Windsor

Mahmoud Khalil
School of Computer Science,
University of Windsor

Alioune Ngom
School of Computer Science,
University of Windsor

ABSTRACT

Video Large Language Models (VideoLLMs) are increasingly deployed for video search and information extraction, where temporally grounded facts must be retrieved from long videos. A key failure mode in this setting is *temporal hallucination*: extracted statements that misattribute when an event occurs, how long it lasts, or what happens before/after, even when relevant evidence exists in the video. Existing hallucination benchmarks largely target free-form video QA or rely on dense temporal supervision, leaving retrieval-style temporal reliability underexplored. **TempHalluc-Bench** is introduced as a *novel* benchmark protocol for evaluating temporal hallucination in retrieval-oriented VideoLLM pipelines, instantiated from **ActivityNet Captions**. **TempHalluc-Bench** is enabled by an annotation-free, post-generation verifier that treats the VideoLLM as a black box and uses only (V, R) at inference. The verifier decomposes an extraction into atomic temporal claims, estimates soft temporal support over time using frozen vision–language encoders, and quantifies inconsistency via mismatch to a text-implied temporal prior. Across diverse VideoLLMs, **TempHalluc-Bench** improves over text-only and similarity-based baselines (Accuracy/F1) and yields stronger accuracy-based reliability signals than prior benchmarks on overlapping models, with gains of up to *+48.2 points Accuracy* (e.g., VideoChatGPT: 78.4 vs. 30.17).

General Terms

Machine Learning, Vision, VideoLLM

Keywords

VideoLLM, temporal hallucination, video information extraction, video search and retrieval, annotation-free verification, long-video understanding

1. INTRODUCTION

Video Large Language Models (VideoLLMs) have recently demonstrated strong performance on complex video understanding tasks and are increasingly deployed as core components of *video search, analytics, and information extraction* systems. In these retrieval-oriented settings, a VideoLLM is not primarily judged by conversational helpfulness, but by its ability to *retrieve and report temporally grounded facts* from long videos—for example, *when* an event occurs, *how long* it lasts, and *what* happens immediately before or after. As real-world deployments shift toward long and unstructured videos, temporal reliability becomes a first-order requirement.

Despite their fluency, VideoLLMs often produce temporally inconsistent extractions. They may attribute an event to the wrong time, assert unsupported persistence (e.g., "throughout the video"), or misorder retrieved events. Such errors, commonly referred to as *temporal hallucinations*, are especially damaging in search-and-extract pipelines: even if the relevant evidence exists somewhere in the video, an incorrect temporal commitment can invalidate the extracted fact and mislead downstream users and systems.

Most existing work evaluates hallucination in VideoLLMs using benchmarks designed for free-form question answering or captioning, frequently relying on dense captions, explicit event boundaries, or manually constructed question–answer pairs. While valuable for controlled diagnosis, these settings do not directly reflect retrieval-oriented use cases, where models are expected to produce short factual extractions with implicit temporal commitments. Moreover, evaluation protocols that require dense temporal supervision are costly to scale and poorly matched to deployment, where such annotations are unavailable.

This work studies temporal hallucination *specifically* in VideoLLM-based video search and information extraction and addresses the following practical question: *can temporal hallucination be evaluated post hoc using only the video and the model's extracted text?* A key observation is that hallucinated extractions tend to induce a measurable mismatch between (i) the temporal structure implied by the generated text (e.g., "early", "after", "throughout") and (ii) the distribution of visual evidence in the video that supports the corresponding claim. This suggests a scalable evaluation strategy that does not require captions or event boundaries at inference time.

Building on this observation, **TempHalluc-Bench** is introduced as a benchmark protocol tailored to temporal reliability in retrieval-oriented VideoLLM pipelines and instantiated from **ActivityNet Captions** [6]. **TempHalluc-Bench** is enabled by a

lightweight, model-agnostic *post-generation verifier* that treats the VideoLLM as a black box and requires only (V, R) at inference. The verifier decomposes an extraction into atomic temporal claims, estimates each claim's *soft* temporal support over the video timeline using frozen vision–language encoders, and measures temporal inconsistency by comparing this support to a text-implied temporal prior. In addition to being annotation-efficient to instantiate, **TempHalluc-Bench** yields *stronger accuracy-based temporal reliability signals* than prior benchmarks on overlapping models—improving the best reported published accuracy by up to *+48.2 points* (e.g., VideoChatGPT: 78.4 vs. 30.17; Table 1)—while remaining consistent with established reliability trends.

The main contributions of this work are:

Novel benchmark protocol for retrieval-style temporal reliability: TempHalluc-Bench is introduced as the first protocol explicitly targeting temporal hallucination in *VideoLLM-based video search and information extraction*, where temporal commitments are central to correctness.

Annotation-free verification at inference: Temporal hallucination evaluation is formulated as a post-generation verification problem, and a principled, model-agnostic verifier is proposed that operates without captions or event boundaries, requiring only (V, R).

Stronger accuracy signals and empirical validation: Experiments across multiple VideoLLMs and long-video datasets demonstrate robust temporal hallucination measurement, trend alignment with recent hallucination benchmarks, and substantially higher accuracy-based signals on overlapping models (up to *+48.2* points; Table 1).

2. RELATED WORK

VideoLLMs are increasingly used as back-end engines for *video search and information extraction*, where the goal is to retrieve *temporally grounded* facts from long videos rather than generate open-ended descriptions. Recent retrieval-augmented and agentic pipelines explicitly frame video understanding as *search-then-summarize/extract* over long-form content, making temporal correctness a first-order requirement for practical deployments [4, 12, 18]. In this setting, *temporal hallucination* manifests as incorrect temporal attribution in extracted facts (e.g., wrong timing, duration, or ordering), which directly degrades retrieval reliability. Related work is reviewed from three perspectives: hallucination in vision–language models, temporal reasoning in video understanding, and post-hoc evaluation/verification of model outputs, highlighting why existing efforts do not fully address temporal hallucination in retrieval-oriented VideoLLM applications.

2.1 Hallucination in Vision–Language Models

Hallucination has been extensively studied in large language models and vision–language models, where systems generate fluent but unsupported content [14, 5]. In video, hallucinations are amplified by long temporal context, dynamic scenes, and evolving object states, leading to errors in event existence, state persistence, and temporal relations. Recent video-focused hallucination benchmarks diagnose these failure modes primarily in QA/caption-style settings, including temporal hallucination and event hallucination evaluations [9, 17, 7, 19]. While these studies provide valuable diagnostics, they are not tailored to *retrieval-style extraction*, where models must output compact factual statements with implicit temporal commitments (e.g., "it happens near the end"), and where correctness depends critically on *when* the retrieved evidence occurs.

2.2 Temporal Reasoning in Video Understanding

Temporal reasoning is a core challenge in video understanding, encompassing action ordering, duration estimation, and temporal localization [2, 8]. Numerous approaches improve temporal modeling via hierarchical representations, temporal attention, and memory-based architectures, often evaluated on supervised localization or long-context comprehension benchmarks [15, 10, 16]. These methods primarily optimize task performance under ground-truth temporal supervision (e.g., labeled moments or segment boundaries). In contrast, the goal is *not* to improve temporal reasoning performance, but to *evaluate temporal hallucination* in VideoLLM search-and-extract pipelines, where dense temporal labels are typically unavailable at inference time.

2.3 Evaluation and Verification of Model Outputs

Post-hoc verification and self-consistency checking have been explored to assess the reliability of generated outputs in language and multimodal systems [13, 1]. In vision–language evaluation, polling- and probing-style protocols have also been proposed to more stably measure hallucinations (e.g., object hallucination) without over-reliance on generation style [11]. More recently, video-oriented reliability estimation has been studied in QA settings using uncertainty- and perturbation-based signals [3]. This work differs in scope and setting: temporal hallucination is targeted in *retrieval-oriented* video search and information extraction, and an evaluation protocol is introduced that is enabled by an annotation-free verifier operating without ground-truth temporal annotations or generator internals, requiring only (V, R) at inference.

3. TEMPHALLUC-BENCH: BENCHMARKING TEMPORAL RELIABILITY FOR VIDEO SEARCH & EXTRACTION

TempHalluc-Bench targets a deployment-relevant regime where a VideoLLM is used as a *search-and-extract* engine: it must find relevant evidence in a long video and return short, factual text that implicitly commits to temporal claims (e.g., when an event happens, what happens before/after, or whether a state persists). Unlike free-form video QA, the goal is not creative description, but reliable extraction of temporally grounded facts.

3.1 Task Setting

Each benchmark sample consists of an input video V and a system output R generated by a VideoLLM conditioned on V (and optionally a user query, which is treated as part of the generation context). The evaluation checks whether R contains *any* temporally unsupported claim with a binary label $y \in \{0, 1\}$. When available, claim-level labels are also supported for finer diagnosis.

3.2 Temporal Hallucination Taxonomy for Extraction

TempHalluc-Bench focuses on temporal errors that frequently appear in retrieval/extraction outputs: (i) **mislocalized timing** (claim is true but at the wrong time), (ii) **ordering errors** (incorrect before/after relations), (iii) **unsupported persistence** (claim incorrectly implies a state holds throughout a long span),

(iv) **unsupported co-occurrence** (events asserted as simultaneous when they are separated in time), and (v) **duration/extent errors** (incorrectly stating how long an event lasts or how many occurrences exist). This taxonomy is intended to be simple enough to annotate at scale, while still covering the temporal reliability questions that matter for search systems.

3.3 Benchmark Construction

TempHalluc-Bench is designed as a *protocol* that can be instantiated on any long-video source. A typical instantiation proceeds as follows:

Video pool: select long videos from a target domain (e.g., instructional, egocentric, surveillance, meetings).

Query set: create time-sensitive prompts that encourage temporal commitments *or* derive them from existing supervision (e.g., rewriting ActivityNet event captions into retrieval intents such as "Find when X happens").

System outputs: run the target VideoLLM(s) to obtain extraction-style outputs R (1–3 sentences or a small list of facts).

Claim decomposition: split R into atomic temporal claims $\mathcal{C} = \{c_i\}_{i=1}^{N}$ (actions, events, or states with implied temporal references).

The benchmark does not require the VideoLLM to output timestamps; instead it evaluates whether the *textual* extraction is temporally supported by the video.

3.4 Annotation Protocol

To obtain gold labels for evaluation, annotators view the video (optionally with coarse navigation aids such as a timeline scrubber) and assign: (i) a response-level label y indicating whether any claim in R is temporally unsupported, and (ii) optional claim-level labels indicating which c_i are problematic. Crucially, dense temporal boundaries or full captions are not required; the annotation burden is closer to verifying extracted statements than to exhaustively segmenting the video. When instantiating from datasets with temporally localized events (e.g., ActivityNet Captions), segment metadata can be used as a navigation aid, but dense boundary annotation is not required during labeling.

3.5 Evaluation Outputs

TempHalluc-Bench supports two complementary evaluation outputs:

Benchmarking generators: aggregate temporal reliability across a model's extractions to characterize generator behavior (e.g., temporal hallucination rate, accuracy of extracted responses, and breakdowns by claim type such as ordering vs. persistence) and form a leaderboard.

Benchmarking verifiers: evaluate a post-hoc detector (e.g., the proposed verifier in §4) against gold labels as a **binary verification** task. Following recent hallucination benchmarks, **Accuracy** is reported as the primary metric and **F1** as a complementary metric robust to class imbalance; the detector commits to a discrete decision using a single threshold selected on validation.

The verifier may also output a continuous hallucination likelihood for analysis, but leaderboard reporting is based on thresholded Accuracy/F1 for comparability.

This work focuses on the *verifier* setting: post-hoc temporal hallucination verification using only (V, R), enabling scalable evaluation when dense temporal supervision is unavailable at inference time.

Algorithm 1 Annotation-Free Temporal Hallucination Verification

Require: Video $V = \{v_t\}_{t=1}^{T}$, extracted response R
Require: Frozen encoders $\phi(\cdot)$, $\psi(\cdot)$; classifier weights $\mathbf{w}$
Ensure: Hallucination score $p(y{=}1 \mid V, R)$
$\mathcal{C} \leftarrow \textsc{ExtractClaims}(R)$ ▷ $\mathcal{C} = \{c_i\}_{i=1}^{N}$
$p^\star \leftarrow 0$
for $c_i \in \mathcal{C}$ **do**
  $a_i(t) \leftarrow \phi(c_i)^\top \psi(v_t)$ ▷ $t = 1, \ldots, T$
  $s_i(t) \leftarrow \mathrm{softmax}_t(a_i(t))$
  $C_i \leftarrow \sum_t s_i(t)^2$; $H_i \leftarrow -\sum_t s_i(t) \log s_i(t)$
  $q_i(t) \leftarrow \textsc{TemporalPrior}(c_i)$; $D_i \leftarrow \mathrm{KL}(q_i \,\|\, s_i)$
  $p_i \leftarrow \sigma(\mathbf{w}^\top [C_i, H_i, D_i])$
  $p^\star \leftarrow \max(p^\star, p_i)$
end for
return $p(y{=}1 \mid V, R) \leftarrow p^\star$

4. METHODOLOGY

An annotation-free verification framework is proposed for evaluating temporal hallucination in VideoLLM-based video search and information extraction systems. Given a video V and a model-generated textual extraction R, the method estimates whether the extracted information is temporally supported by the visual content. The framework operates entirely post-generation and treats the VideoLLM as a black box, requiring no ground-truth captions, event boundaries, or model-specific supervision. This design is suitable for benchmarking temporal reliability in real-world retrieval settings, where only (V, R) are available.

4.1 Overview

The verifier decomposes the extracted response into atomic temporal claims, estimates each claim's *soft* temporal support over the video timeline, measures temporal inconsistency between the claim text and the visual evidence, and aggregates claim-level scores into a response-level hallucination estimate. Figure 1 summarizes the pipeline and Algorithm 1 gives the full procedure.

4.2 Problem Setup

Let $V = \{v_t\}_{t=1}^{T}$ denote a video represented as a sequence of sampled frames or segments, and let R denote a textual output generated by a VideoLLM in response to a search or information extraction query. Unlike free-form dialogue, R is assumed to contain factual statements intended to retrieve temporally grounded information from V.

R is decompose into a set of N atomic temporal claims $\mathcal{C} = \{c_i\}_{i=1}^{N}$, where each claim corresponds to an asserted action, event, or object state with an implied temporal reference. The objective is to estimate whether any claim in $\mathcal{C}$ is temporally hallucinated, i.e., not supported by the visual evidence at the implied time. This is formalized as predicting a binary variable $y \in \{0, 1\}$ indicating whether the extracted response R contains temporal hallucination.

4.3 Soft Temporal Support

For each claim c_i, a temporal support distribution is estimated over the video timeline:

$$s_i(t) \triangleq p(t \mid c_i, V), \quad t \in \{1, \ldots, T\}. \tag{1}$$

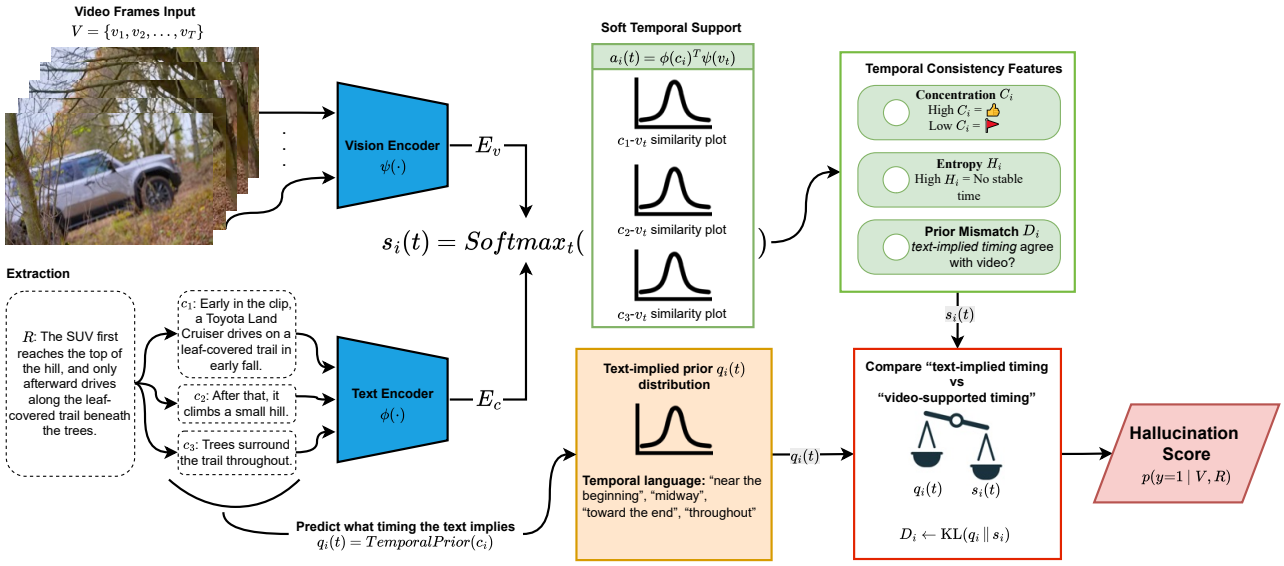

Fig. 1: Annotation-free temporal hallucination verification for retrieval-oriented VideoLLM outputs. Given (V, R), R is decomposed into claims, estimate soft temporal support $s_i(t)$ via frozen vision–language encoders, compare it to a text-implied prior $q_i(t)$, and aggregate claim scores to obtain a response-level hallucination likelihood.

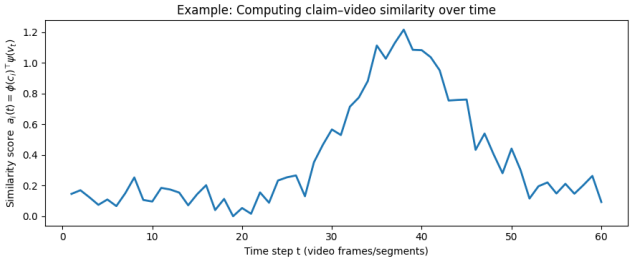

Fig. 2: Plot of $a_i(t)$, the similarity score over time for a single claim versus each video segment.

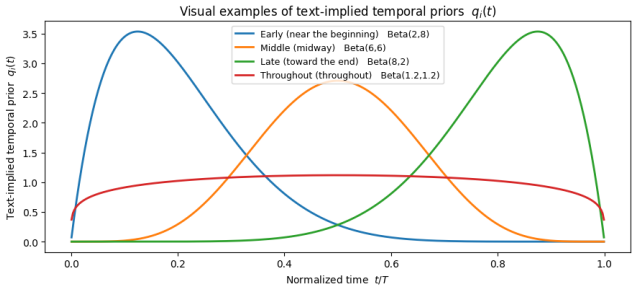

Fig. 3: Visual example of the text-implied temporal prior $q_i(t)$ — i.e., what timing the text alone suggests, before looking at the video.
- "near the beginning" → a distribution peaked early
- "midway" → peaked in the middle
- "toward the end" → peaked late
- "throughout" → relatively spread across the whole timeline

This grounding is not directly observed and is inferred from cross-modal alignment. The claim is encoded with a frozen text encoder $\phi(\cdot)$ and each video segment with a frozen visual encoder $\psi(\cdot)$:

$$a_i(t) = \phi(c_i)^\top \psi(v_t), \qquad s_i(t) = \frac{\exp(a_i(t))}{\sum_{t'=1}^{T} \exp(a_i(t'))}. \tag{2}$$

The resulting $s_i(t)$ is a continuous estimate of visual support over time, avoiding brittle thresholding. Figure 2 shows an example of computing $a_i(t)$, claim-video similarity over time.

4.4 Temporal Consistency Features

Temporal hallucination in retrieval/extraction outputs often manifests as a mismatch between the temporal structure implied by a claim and the distribution of visual evidence. This is captured with three complementary, continuous features derived from $s_i(t)$.

4.4.0.1 *Concentration..*

$$C_i = \sum_{t=1}^{T} s_i(t)^2. \tag{3}$$

4.4.0.2 *Entropy..*

$$H_i = -\sum_{t=1}^{T} s_i(t) \log s_i(t). \tag{4}$$

4.4.0.3 *Text-implied prior and mismatch..*

Claim text often implies a temporal scope or bias (e.g., "before", "after", "throughout"). This is modeled with a smooth prior $q_i(t \mid c_i)$ over normalized time $t/T \in (0, 1)$, parameterized as a Beta distribution whose parameters are predicted from c_i by a small neural module.

Figure 3 shows a visual example of the text-implied temporal prior $q_i(t)$.

4.5 Compare "text-implied timing" vs "video-supported timing"

Grounding–prior mismatch is then measured as:

$$D_i = \mathrm{KL}\big(q_i(t \mid c_i) \,\|\, s_i(t)\big). \tag{5}$$

Meaning if the text says "near the beginning" (prior q_i peaks early), but the video evidence peaks late (s_i peaks late), then D_i becomes large, hence strong sign of temporal hallucination. This is the key "consistency check". Figure 4 shows an example of comparing "text-implied timing" $q_i(t)$ vs. "video-supported timing" $s_i(t)$.

4.6 Hallucination Scoring

For each claim c_i, $\mathbf{f}_i = [C_i, H_i, D_i]$ is formed and a claim-level hallucination probability is estimated:

$$p(y_i{=}1 \mid V, c_i) = \sigma(\mathbf{w}^\top \mathbf{f}_i). \tag{6}$$

Aggregation to the response level is performed by max-pooling over claims:

$$p(y{=}1 \mid V, R) = \max_i p(y_i{=}1 \mid V, c_i), \tag{7}$$

reflecting the retrieval-oriented assumption that any temporally incorrect fact compromises the extracted response.

4.7 Training and Inference

The classifier weights $\mathbf{w}$ (and the small prior module) are trained using labeled examples from benchmark data, while all encoders remain frozen. At inference, the verifier requires only (V, R) and does not access captions, temporal annotations, or generator internals, enabling scalable temporal reliability benchmarking in long-video retrieval pipelines.

4.8 Theoretical Intuition

Temporal hallucinations arise when a response encodes a temporal structure that is not supported by the underlying video. From a probabilistic perspective, this corresponds to a mismatch between a claim's latent temporal support distribution $s_i(t)$ and the temporal prior implied by the claim text $q_i(t \mid c_i)$. When a claim is temporally correct, visual support tends to be concentrated and aligned with the bias suggested by the text; hallucinated claims typically exhibit diffuse or unstable support across time. The proposed method operationalizes this intuition by representing grounding as a distribution over time and measuring consistency via concentration/entropy and the divergence D_i, enabling detection without dense temporal supervision.

5. EXPERIMENTS

TempHalluc-Bench is evaluated as a temporal reliability benchmark for retrieval/extraction-style VideoLLM outputs: given a video V and a model-generated extraction R, the task is to determine whether R contains *any* temporally unsupported claim. The experimental design follows reporting conventions of recent video hallucination benchmarks (binary judgment with accuracy as the primary metric) [9, 17, 7, 19]. Results are organized around (i) datasets, protocol, and metrics, (ii) main verifier performance, (iii) generalization across VideoLLMs, (iv) comparison to established temporal hallucination benchmarks, and (v) discussion (robustness, ablations, efficiency, qualitative analysis).

Research Questions.

RQ1: Can temporal hallucinations be detected *post hoc* using only (V, R) (no captions/temporal annotations at inference)?

RQ2: Does the verifier generalize across different VideoLLMs used for search/extraction?

RQ3: How sensitive is performance to encoder choice and temporal sampling?

RQ4: Do model reliability trends agree with established temporal hallucination benchmarks?

5.1 Datasets, Protocol, and Metrics

*5.1.0.1 **TempHalluc-Bench** (retrieval/extraction setting; instantiated from ActivityNet Captions)..* **TempHalluc-Bench** is instantiated from **ActivityNet Captions** [6], an untrimmed long-video dataset with temporally localized event segments and natural-language descriptions. For each ActivityNet video, an event description is treated as a retrieval-style query and a VideoLLM is prompted to produce a short extraction R that reports the relevant fact(s) together with an implicit or explicit temporal attribution (e.g., "early", "after X", "near the end", or an approximate time). Each sample consists of a video V and a model-produced extraction R (answer snippet / extracted facts / short retrieval summary). The ground-truth label $y \in \{0, 1\}$ indicates whether R contains *at least one* temporally unsupported claim with respect to the underlying video evidence. When claim-level annotations are available, evaluation is also performed at the claim level using $\mathcal{C} = \{c_i\}_{i=1}^{N}$.
Concretely, each ground-truth event caption is converted into a retrieval query by lightly rewriting it into a *search intent* (e.g., removing descriptive modifiers and phrasing it as "Find when event happens"), and the VideoLLM is prompted to return a short extracted answer with an explicit or implicit temporal attribution.

5.1.0.2 Train/val/test split and supervision.. **TempHalluc-Bench** labels are used as the *only* source of supervision for training the verifier. **ActivityNet videos** are split into **70/10/20** train/validation/test (by video, to prevent leakage across splits). Only the lightweight verifier components (classifier weights $\mathbf{w}$ and the temporal-prior module) are trained on the **train** split, a single operating threshold is selected on the **validation** split (maximizing F1), and all metrics are reported on the held-out **test** split. Encoders remain frozen throughout. Unless explicitly stated (e.g., model-holdout evaluation in Table 3), no external benchmark labels are used for training or tuning.

5.1.0.3 Comparison benchmarks (trend alignment only).. To contextualize model-level trends (RQ4), comparisons are made against published temporal reliability signals from recent hallucination benchmarks: **VIDHALLUC** [9], **EventHallusion** [17], **NOAH** [7], and **VERHallu** [19]. These benchmarks are used only for *trend alignment* and cross-reporting on overlapping models; no training is performed on them.

5.1.0.4 Metrics.. Following dominant reporting practice in hallucination benchmarks (binary decision $\rightarrow$ accuracy) [17, 7], temporal hallucination verification is treated as a **binary classification** problem and the following metrics are reported:

Accuracy (%) as the **primary** metric.

F1 score as a complementary metric, robust to class imbalance.

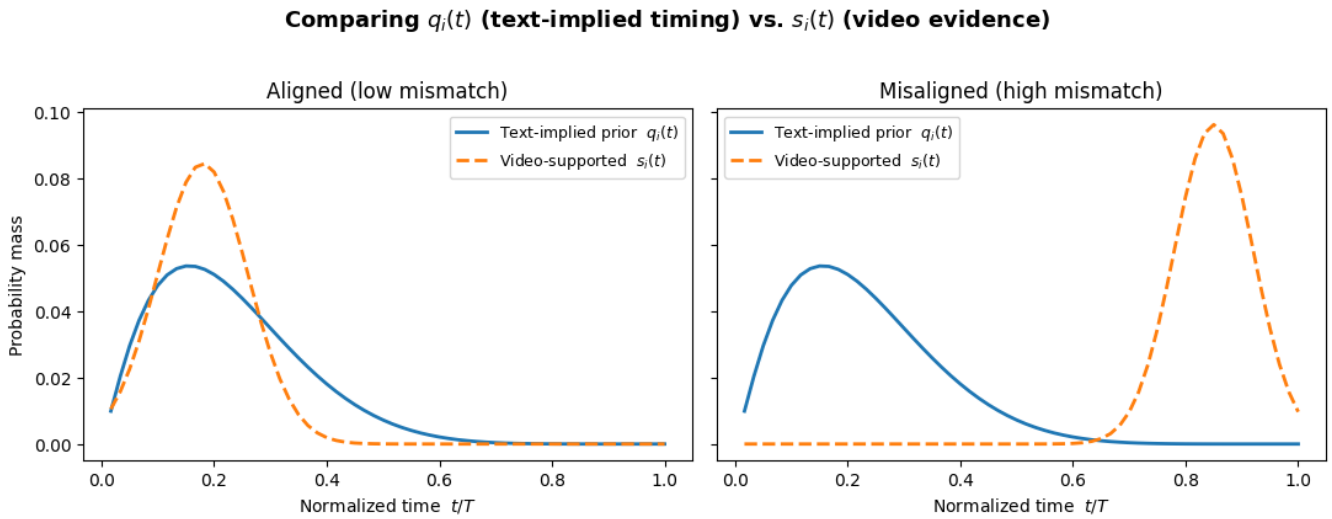

Fig. 4: An example of comparing "text-implied timing" $q_i(t)$ vs. "video-supported timing" $s_i(t)$:
- Left (Aligned / low mismatch): the text implies the event happens early (blue $q_i(t)$) and the video evidence also peaks early (dashed $s_i(t)$) $\rightarrow$ low KL divergence.
- Right (Misaligned / high mismatch): the text implies the event happens early (blue), but the video evidence peaks late (dashed). $\rightarrow$ high KL divergence, strong signal of temporal hallucination.

Model	VIDHALLUC	EventHallusion	NOAH TempQA	TempHalluc-Bench
VideoChatGPT	30.17	24.21	–	78.4
Video-LLaVA	27.17	46.32	39.0	81.2
LLaVA-NeXT-Video	–	69.93	10.0	83.7
VILA (VILA1.5)	63.33	65.04	–	85.1
Qwen2.5-VL 72B	–	–	68.2	86.8
GPT-4o	–	84.11	14.8	**90.5**

Table 1. : **Cross-benchmark accuracy comparison.** Published metrics are from **VIDHALLUC** [9], **EventHallusion** [17], and **NOAH** [7]. The rightmost column reports **TempHalluc-Bench** (ours) verification accuracy for the same models in the retrieval/extraction setting. The bold value indicates the highest accuracy in the table. **Note:** "–" indicates the corresponding benchmark did not report results for that model/metric.

Thresholding. The verifier outputs a continuous hallucination likelihood $h(V, R) \in [0, 1]$; a single operating threshold is selected on the validation split (maximizing F1) and Acc/F1 are reported on the test split.

5.2 Evaluated VideoLLMs

The verifier is evaluated on retrieval/extraction outputs produced by a diverse set of VideoLLMs, prioritizing models commonly reported in temporal hallucination benchmarks for comparability—including VideoChatGPT, Video-LLaVA, LLaVA-NeXT-Video, and VILA (VILA1.5) [9, 17]—and adding strong proprietary baselines when available (e.g., GPT-4o). Unless otherwise noted, the verifier treats all VideoLLMs as black boxes and operates on their generated extractions R.

5.3 Encoders, Sampling, and Implementation Details

5.3.0.1 Default encoders.. To keep the benchmark stable and reviewer-friendly, the proposed default instantiation uses widely adopted frozen encoders:

Visual encoder ψ: CLIP ViT-L/14 (frozen).

Text encoder ϕ: matched CLIP text encoder (frozen).

5.3.0.2 Temporal sampling.. Unless stated otherwise, T frames/segments are uniformly sampled over the video and each

Method	Acc. (%)	F1
Random	50.0	50.0
Text-only	56.4	55.2
Global Similarity	59.2	58.1
Max Similarity	60.5	59.4
Entropy-only	63.9	62.6
Ours (Temporal Consistency)	**70.8**	**69.7**

Table 2. : Response-level temporal hallucination verification on **TempHalluc-Bench**. Accuracy and F1 are reported using a single threshold selected on the validation split.

VideoLLM (unseen at train)	Acc. (%)	F1
VideoChatGPT	78.4	68.0
LLaVA-NeXT-Video	83.7	69.9
GPT-4o	90.5	71.1

Table 3. : **Cross-model generalization.** Trained on **Video-LLaVA** and **VILA (VILA1.5)**; evaluated on unseen VideoLLMs.

is encoded independently. Robustness to temporal resolution is evaluated with $T \in 8, 16, 32, 64$ in §*6.0.3*.

5.3.0.3 Training protocol.. Only the lightweight verifier components ($\mathbf{w}$ and the temporal-prior module) are trained on **TempHalluc-Bench** labels; all encoders remain frozen. At inference, the verifier requires only (V, R).

5.4 Baselines

Simple baselines commonly expected in hallucination verification are used for comparison [9, 17]:

Random: uniform prediction.

Text-only: classifier over text embeddings of R (no video).

Global similarity: average similarity between R and the video over time.

Max similarity: maximum frame similarity over time.

Entropy-only: uses temporal entropy aggregated over claims (uncertainty-only signal).

The full method combines temporal support, uncertainty, and grounding–prior mismatch (§4)

5.5 Main Results: Temporal Hallucination Verification

Table 2 reports response-level verification on **TempHalluc-Bench**. The proposed verifier substantially improves over text-only, similarity-based, and uncertainty-only baselines, supporting **RQ1**. Gains are most pronounced when extractions commit to ordering or temporal scope (e.g., "before/after", "throughout").

5.6 Generalization Across VideoLLMs

To test **RQ2**, a model-holdout protocol is adopted: the verifier is trained on extraction outputs from **Video-LLaVA** and **VILA (VILA1.5)** and evaluated on unseen generators (**VideoChatGPT**, **LLaVA-NeXT-Video**, and **GPT-4o**). Table 3 indicates that strong performance is maintained on held-out generators, suggesting that model-agnostic temporal inconsistency patterns are captured rather than overfitting to a single VideoLLM.

Variant	Acc. (%)	F1
Concentration only (C)	63.9	62.5
$C + H$	67.1	65.8
$C + D$	68.5	67.2
$C + H + D$ **(full)**	**70.8**	**69.7**

Table 4. : Ablation of temporal consistency signals.

Encoder (vision/text)	Acc. (%)	F1
CLIP ViT-B/32	70.8	69.7
CLIP ViT-L/14	71.3	70.2

Table 5. : Sensitivity to encoder scale (matched text encoder).

5.7 Comparison to Established Temporal Hallucination Benchmarks

A key goal is that **TempHalluc-Bench** is *not isolated*: temporal reliability trends should be consistent with established benchmarks that probe temporal errors (**RQ4**). Table 1 summarizes published accuracy-based temporal hallucination signals from **VIDHALLUC**, **EventHallusion**, and **NOAH**, and reports **TempHalluc-Bench** verification accuracy in the rightmost column for direct comparison ("–" indicates the corresponding benchmark did not report results for that model/metric). On overlapping models, **TempHalluc-Bench** provides substantially stronger accuracy-based temporal reliability signals while preserving consistent ranking trends.

6. DISCUSSION AND ABLATION

6.0.1 Ablation Study. Temporal consistency components are ablated to quantify their contributions. Table 4 shows that each component improves verification quality, with the largest gains obtained by incorporating the text-implied temporal prior mismatch D.

6.0.2 Encoder Sensitivity. To show the benchmark is not "encoder-specific", common CLIP scales are evaluated. Results in Table 5 show modest variation.

6.0.3 Temporal Sampling Robustness. Temporal sampling affects temporal reasoning and hallucination behavior [17, 7]. The number of sampled segments T is varied while keeping the verifier fixed, and Acc/F1 are reported. In practice, T=32 provides a strong accuracy–cost tradeoff for long videos; reporting a small curve or a compact table in the appendix is recommended.

6.0.4 Efficiency. Wall-clock time per video is reported for (i) feature extraction (encoder forward pass) and (ii) scoring, along with throughput (videos/sec) on a single GPU. This "cost" reporting mirrors prior benchmark work and strengthens deployment realism [9].

6.0.5 Qualitative Analysis. Representative examples (correct, hallucinated, borderline, failure) are included with: (i) extracted claim(s), (ii) temporal support $s_i(t)$, (iii) the overlaid text-implied prior $q_i(t)$, and (iv) a few key frames at the top-k support times. This presentation makes temporal inconsistencies interpretable and aligns with common qualitative "aids" in hallucination/grounding papers.

Overall, **TempHalluc-Bench** supports reliable evaluation of temporal hallucination in retrieval/extraction outputs. Results indicate that temporal hallucinations are detectable using only (V, R), generalize across VideoLLMs, and yield trends consistent with established temporal hallucination benchmarks [17, 7]. These findings motivate benchmarking temporal reliability specifically in the search-and-extract regime, where a single temporally incorrect fact can invalidate downstream use.

7. CONCLUSION AND FUTURE WORK

This paper targets a practical reliability gap in *VideoLLM-based video search and information extraction*: even when relevant evidence exists in a long video, models may misstate *when* it occurs, producing temporally unsupported extracted facts. **TempHalluc-Bench** is introduced as a benchmark protocol for measuring temporal hallucination in retrieval-oriented settings, together with a lightweight, model-agnostic verifier that operates post-generation. The verifier decomposes an extraction into atomic temporal claims, estimates each claim's soft temporal support over the video timeline using frozen vision–language encoders, and aggregates concentration, uncertainty, and grounding–prior mismatch signals into a response-level hallucination likelihood. Experiments across diverse VideoLLMs and long-video datasets show that temporal hallucination in extraction pipelines can be evaluated without access to captions or event boundaries at inference time, and that the resulting reliability trends are consistent with signals reported by recent hallucination benchmarks while providing complementary coverage tailored to search-and-extract use cases.

TempHalluc-Bench also has limitations. The verifier depends on frozen encoders and soft temporal grounding, which may under-capture fine-grained cues in cluttered scenes, subtle state changes, or rapid event transitions. Moreover, while the protocol avoids dense temporal supervision, benchmark instantiation still requires human verification to produce gold labels for auditing generators and for reporting verifier quality.

Future work should improve deployment realism while preserving annotation efficiency. Query and extraction formats can be expanded to better match real video search behavior, including first occurrence, immediate before/after relations, and repeated events, and models can be stress-tested under ambiguous or adversarial queries. Beyond binary outcomes, *graded* temporal reliability can be studied to reflect error severity (e.g., slight mislocalization versus fundamentally incorrect ordering). Evaluation can also be extended from short textual extractions to structured outputs used in practical systems—such as timestamped evidence snippets and multi-claim reports—and retrieval-time verification, calibrated abstention, and feedback mechanisms can be investigated to improve temporal reliability in end-to-end search pipelines.

8. REFERENCES

[1] Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar R. Zaiane. Faithfulness in natural language generation: A survey of methods and metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022.

[2] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[3] Sushant Gautam, Cise Midoglu, Vajira Thambawita, Michael A. Riegler, and Pål Halvorsen. Videohedge: Entropy-based hallucination detection for video-vlms via semantic clustering and spatiotemporal perturbations, 2026.

[4] Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. Videorag: Retrieval-augmented generation over video corpus. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21278–21298, July 2025.

[5] Ziwei Ji, Nayeon Lee, Rita Frieske, Tianyi Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023.

[6] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

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

[8] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Temporal grounding of natural language descriptions in videos. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[9] Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[10] Linjie Li, Jie Lei, Zhe Wang, Jingjing Li, Jason Kuen, Zheng Feng, Dongyu Chen, Jianfei Cai, Mike Zheng Shou, Rui Yan, et al. Hero: Hierarchical encoder for video+language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[11] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[12] Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-rag: Visually-aligned retrieval-augmented long video comprehension. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. NeurIPS 2025 Poster.

[13] Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore, December 2023. Association for Computational Linguistics.

[14] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[15] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video

and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[16] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[17] Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, Xingjun Ma, and Jingjing Chen. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*, 2024.

[18] Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[19] Zefan Zhang, Kehua Zhu, Shijie Jiang, Hongyuan Lu, Shengkai Sun, and Tian Bai. Verhallu: Evaluating and mitigating event relation hallucination in video large language models. *arXiv preprint arXiv:2601.10010*, 2026.