Diagnosing and Correcting Bias in MLLM for Long Video Understanding

Xusheng Liang[1,2] **Jianqiao Sun**[3,4*] **Hao Zhang**[3] **Yulei Niu**
Hengshuang Zhao[5,2] **Jiawei Ma**[1]
[1]City University of Hong Kong [2]Shenzhen Loop Area Institute [3]Xidian University
[4]JX Intelligence Technology Co., Ltd. [5]The University of Hong Kong
xushliang2-c@my.cityu.edu.hk jiaweima@cityu.edu.hk

Abstract

Long-video question answering is challenging since the answer often hinges on a few decisive moments scattered throughout the video, while memory constraints force Multimodal Large Language Models (MLLMs) to sample frames at extremely low rates, resulting in sparsity that risks omitting key evidence. Though aggressively enlarging the context size is intuitive, we argue that performance remains fundamentally constrained by the bias, *i.e.*, the misleading cues caused by spurious *linguistic* correlations and salient yet irrelevant *visual* observations. In this paper, we first introduce S-MME, a benchmark designed to systematically diagnose shortcut-biased behavior in MLLMs. To mitigate this failure, we further propose a training-free framework, Counterfactual Long-video Evidence-Aware Reasoning (CLEAR). By treating the full video-question pair as a complete view, CLEAR estimates bias effects via augmented views by only keeping the question or locally salient visual clips. Then, we intervene on hidden states to mitigate the discrepancy at inference such that the correct answer can be predicted confidently. With experimental studies across five benchmarks, we show that our method is generalizable and can be integrated in different MLLMs, yielding consistent gains in prediction and robustness. We hope CLEAR can offer a complementary direction in test-time scaling for long-video understanding and will publish code.

1 Introduction

Visual understanding over a long temporal horizon is inherently challenging because the correct answer often hinges on a few decisive frames, *i.e.*, *evidence*, sparsely dispersed across the video. Then, memory constraints force current Multimodal Large Language Models (MLLMs) to sample frames at extremely low rates as the input context, further compounding this difficulty and inducing the model to rely on spurious correlations, *i.e.*, *shortcuts*, for prediction. To mitigate predictions, the intuitive methods are to scale up the sampling rate and context window [49], increase visual token capacity [36], and strengthen instruction tuning [6]. Nevertheless, they all incur substantially higher computational cost and inevitably face an efficiency–effectiveness trade-off.

At the same time, enlarging the context alone is insufficient, and performance remains fundamentally constrained by biases even when relevant frames are completely fed into the model. As illustrated in Fig. 1, the model is prone to follow misleading cues, including (left) spurious linguistic correlations [1], where the phrasing of the question and answer options makes a distractor plausible without viewing the video, and (right) salient yet causally irrelevant observations, *i.e.*, frames that are semantically related to the query yet do not determine the answer. Similarly, while existing methods such as frame selection [47], token pruning [31] and memory mechanisms [33], aim to filter out

*Contribution partially done while Jianqiao Sun was an intern at JX.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

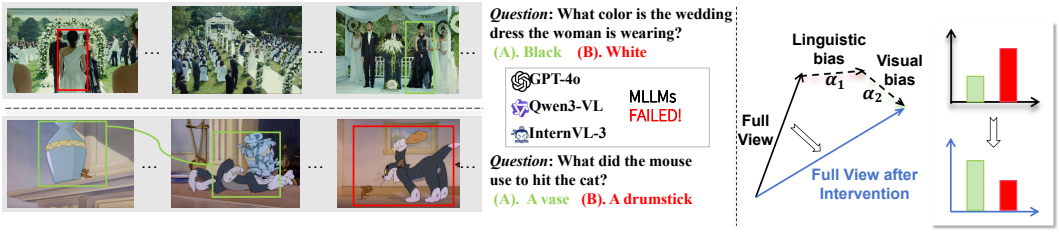


Figure 1: MLLMs often fail despite the presence of decisive evidence, either by following spurious linguistic correlations or by over-relying on salient but misleading visual observations. CLEAR diagnoses these shortcut-biased trajectories with controlled probes and steers the full-video representation toward evidence-consistent prediction.

irrelevant frames before feeding them into the model, they still assume that the model can be robust enough to use the key observation. To this end, we believe that evaluation should go beyond aggregate benchmark scores and systematically target samples that are sensitive to shortcuts.

In detail, we construct S-MME, standing for *Shortcuts in Multi-Modal Evaluation*, a long-video benchmark composed of video-question pairs designed to expose shortcut-sensitive failures, *i.e.*, when given a restricted observation that forces a spurious correlation, models produce an incorrect prediction, whereas access to the evidence frames can yield the correct answer. Then, following the examples in Fig. 1 and adopting the concept of counterfactual samples [27], we construct the restricted view in two aspects: a question-only view where the question and options alone make distracted prediction, and a salient-yet-irrelevant view where individual clips have high semantic similarity to the question in a common representation space yet lack the decisive evidence required for the correct answer. In this way, we can directly test whether a model follows the decisive clips or instead collapses to a counterfactual alternative.

To further mitigate the bias to such spurious shortcut, we propose a training-free framework Counterfactual Long-video Evidence-Aware Reasoning (CLEAR). CLEAR runs the frozen model, treats the full video-question pair as a complete view, and constructs counterfactual views that maximally preserve the shortcut signals and weaken or even remove decisive evidence. Specifically, we use a question-only view to expose linguistic bias and a salient view formed from highly similar visual observations to expose visual bias. We posit that the prediction discrepancy between views serves as an estimate of the bias effect, and we apply lightweight residual steering in deep layers to intervene on hidden states and suppress spurious correlations, thereby minimizing the distraction and enabling the model to predict the correct answer with more confidence. Meanwhile, our method is generalizable and can be integrated into different MLLMs without additional training, and yields consistent gains through comprehensive experimental studies. Finally, our contributions are summarized as follows:

- We identify bias from 1) spurious linguistic correlations and 2) salient but misleading visual observations as a fundamental bottleneck in long-video understanding by MLLMs, which persists beyond sparse frame sampling alone.
- To examine such bias of existing MLLMs, we introduce S-MME, a benchmark for systematically diagnosing this behavior. The S-MME considers spurious correlation caused by either linguistic correlation or salient-yet-misleading views.
- We propose CLEAR, a training-free framework that estimates bias effects from counterfactual views and mitigates them through hidden-state intervention, thereby encouraging predictions to rely on decisive evidence. The method is generalizable and can be applied to different MLLMs operating in either classification or next-token prediction modes.
- Through experimental study, we demonstrate consistent improvements over three public datasets, Video-MME [11], LongVideoBench [41], and MLVU [54], as well as our S-MME benchmark.

2 Related Works

Multimodal LLMs for Long-Video Understanding Recent Video-LLMs improve long-video understanding by increasing the amount of visual evidence that can be processed within limited model context. Existing methods typically compress frame-level features into compact visual

tokens [22, 56, 46, 38, 20], reduce spatial-temporal redundancy through pooling, token aggregation, or adaptive filtering [26, 45, 44, 12], or extend multimodal context length to support more frames and longer reasoning horizons [49, 36, 35, 5, 24]. While these methods improve evidence access, often with non-trivial memory or computational cost [17, 32, 15, 14, 13], our work studies a complementary problem: whether accessible evidence is faithfully used when shortcut cues provide plausible answers.

Shortcut Learning and Bias in Visual Question Answering Shortcut learning has long been recognized as a central challenge in visual question answering. Early studies showed that VQA models can exploit answer priors, question templates, or dataset correlations without sufficient visual grounding [1, 3, 16, 18, 29]. Subsequent work exposed and mitigated such biases through distribution-shifted benchmarks, language-bias suppression, stronger visual grounding, and counterfactual sample construction [1, 10, 7, 27]. While these studies provide important insights, they mainly focus on static-image VQA or conventional multimodal fusion. Long-video reasoning introduces an additional shortcut source: a locally salient clip may appear predictive but miss temporally distributed events, state changes, or later evidence required for the correct answer. Recent video QA benchmarks also suggest that final answer accuracy alone does not fully reveal whether models ground their predictions in the relevant temporal evidence [43, 41, 23]. Our work extends shortcut diagnosis from language-only bias to both language-prior and partial-observation shortcuts in long-video reasoning.

Training-Free Methods for Long-Video Understanding Training-free long-video methods improve frozen Video-LLMs through inference-time input selection, memory construction, or agentic reasoning, without updating model parameters [19]. Selection-based methods reduce redundant visual input through adaptive keyframe sampling, token merging, or query-guided retrieval [34, 50, 47, 52]. Memory-enhanced methods maintain long-range information using event memory, temporal caches, recursive summarization, or tree-structured video representations [33, 9, 30, 40]. Agent-based methods formulate long-video QA as an active inference process, iteratively retrieving segments, revising plans, and performing multi-step reasoning with language agents [37, 4, 53, 39, 42]. These approaches mainly optimize which evidence is selected, retained, or queried during inference.

3 S-MME: Shortcuts in Multi-Modal Evaluation

3.1 Making Evidence Use Diagnosable

Long-video QA evaluation typically reports whether a model produces the correct answer. However, final-answer accuracy alone does not reveal whether the model has relied on the decisive evidence. A model may fail because the decisive evidence is missing from the sampled context, but it may also fail even when such evidence is available, because the prediction is dominated by shortcut cues such as spurious linguistic correlations or salient-yet-misleading visual observations. We refer to the latter as biased evidence use: the model observes relevant evidence but still follows a plausible shortcut. This behavior is difficult to isolate with long-video benchmarks, where a wrong answer may arise from missing frames, temporal ambiguity, or shortcut-induced bias. To make evidence use diagnosable, we construct cases in which complete video evidence and restricted shortcut cues support competing answers. If a model predicts the shortcut-consistent distractor under the full-video input, the failure indicates that the model does not fully use the decisive evidence. Accordingly, we introduce S-MME, a diagnostic benchmark for evaluating shortcut-robust long-video reasoning.

3.2 Shortcut-Sensitive Cases

The basic unit of S-MME is a shortcut-sensitive long-video QA case. Each case consists of a long video, a question, answer candidates, a sufficient evidence span, and an annotated shortcut-consistent distractor. A valid conflict case must satisfy three conditions. First, the full video contains sufficient evidence to determine a unique correct answer. Second, at least one restricted observation, such as the question alone or a short salient clip, is insufficient for determining the correct answer. Third, this restricted view makes a specific incorrect answer plausible, yielding a shortcut-consistent distractor.

We consider two dominant shortcut sources. The first comes from spurious linguistic correlations, where the question or answer options alone make an incorrect answer plausible without visual grounding. The second comes from salient-yet-misleading visual observations, where a visually salient segment is semantically related to the question but does not determine the answer without

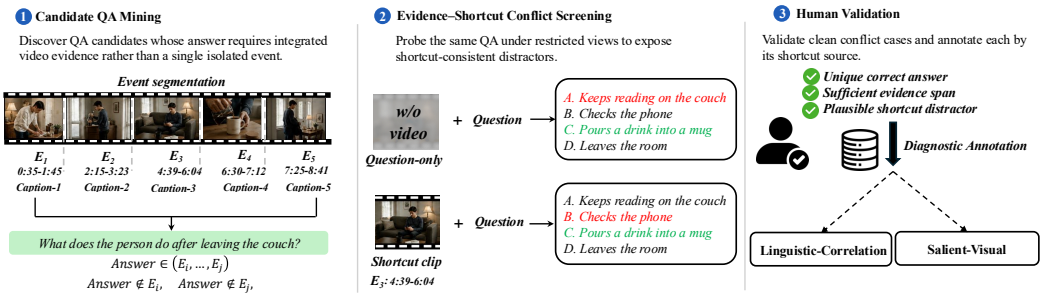


Figure 2: **Construction pipeline of S-MME.** We first mine candidate questions with explicit long-video evidence requirements, then use restricted views to identify shortcut-consistent distractors, followed by human validation of answer uniqueness, evidence sufficiency, and shortcut plausibility.

Table 1: Statistics of the S-MME diagnostic benchmark.

Category	Count	Avg. Dur. (min)	Min Dur. (min)	Max Dur. (min)
Linguistic-Correlation Conflict	148	11.9	5.6	74.7
Salient-Visual Conflict	172	17.6	12.8	106.6
Total	**320**	**15.0**	**5.6**	**106.6**

broader temporal context. These two sources reflect common shortcut pathways in long-video QA and provide counterfactual views for diagnosis. Beyond assigning each case to one shortcut source, S-MME annotates the sufficient evidence span and the shortcut-consistent distractor for every sample.

3.3 Benchmark Construction

We construct S-MME through a three-stage pipeline, as illustrated in Figure 2. The first stage mines candidate QA samples with explicit long-video evidence requirements. Given a long video, we segment it into coherent temporal events and summarize the timeline with concise event descriptions. Candidate questions are generated from these event summaries to require evidence beyond a single isolated observation, such as cross-event comparison, temporal state change, delayed outcome, or local observation contradicted by later evidence. For each candidate, annotators mark the sufficient evidence span and verify that the full video uniquely determines the correct answer.

The second stage uses restricted views to identify shortcut-consistent distractors. For the question-only view, annotators check whether the question and answer options alone make a particular incorrect option plausible without visual grounding. For the salient-view, annotators identify a short segment that is semantically related to the question and supports a plausible distractor, but does not contain sufficient evidence to determine the correct answer without broader video context. These restricted views are used only for candidate discovery and do not determine the ground truth of a sample.

The third stage applies human validation and quality control. Annotators retain a candidate only if it satisfies three criteria: (i) the full video uniquely supports the correct answer, (ii) the restricted view is insufficient for determining the correct answer, and (iii) the annotated shortcut-consistent distractor is plausible under the restricted view. Ambiguous questions, multiple-answer cases, noisy videos, and samples without a clearly identifiable evidence span are removed. This meticulous validation pipeline ensures that S-MME evaluates whether models can leverage available evidence to override plausible shortcuts, effectively decoupling reasoning robustness from annotation noise or missing information.

4 CLEAR: Counterfactual Long-video Evidence-Aware Reasoning

4.1 Problem Setup

We consider a video MLLM parameterized by θ, which takes a video v and a textual query x as input, and generates a response $y = (y_1, \ldots, y_T)$ in an auto-regressive manner:

$$y_t \sim p_\theta(y_t \mid v, x, y_{<t}), \tag{1}$$

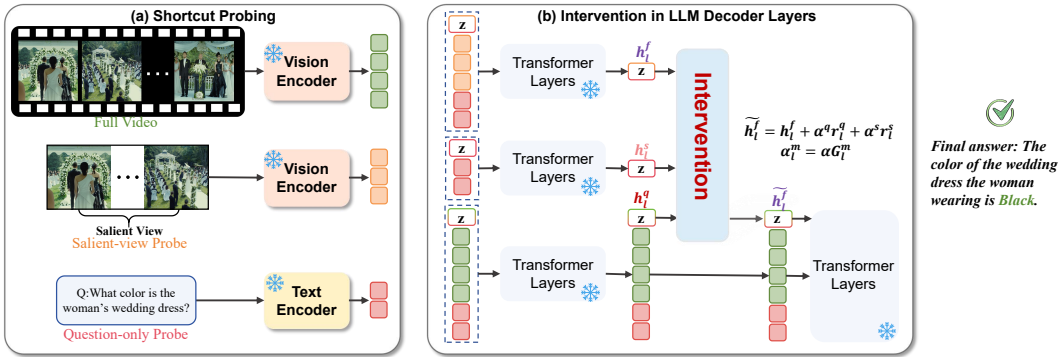


Figure 3: **Overview of CLEAR.** (a) The full-video path produces the final answer, while question-only and salient-view probes reveal spurious linguistic correlations and salient-yet-misleading visual observations. The salient view is formed by selecting top-K locally relevant clips according to query-conditioned visual attention. (b) CLEAR extracts the hidden state of the shared decision token z at selected late decoder layers, and constructs residual vectors $r_l^q = h_l^f - h_l^q$ and $r_l^s = h_l^f - h_l^s$ from the restricted-view probes to the full-video state. These residual vectors are injected into the full-video path to steer its hidden state away from shortcut anchors before final decoding.

where $y_{<t}$ denotes previously generated tokens.

Long-video understanding requires the model to integrate evidence across temporally distant clips. Errors may still occur when the relevant evidence is present, because the model can form a reasoning trajectory dominated by biases. CLEAR addresses this issue by exposing such shortcut trajectories through counterfactual views and correcting the full-video trajectory through residual steering.

4.2 Shortcut Probing via Counterfactual Views

To make shortcut-biased reasoning observable, CLEAR evaluates the same video-question sample under counterfactual observations. These observations are not used to produce the final answer. Instead, they provide diagnostic trajectories that reveal how the model behaves when complete video evidence is removed or restricted. We focus on two common shortcut sources in long-video reasoning: language priors from the question itself and local visual shortcuts from salient but incomplete clips.

Question-only probe. The question-only probe removes the video input and keeps only the textual query. This view captures the model's spurious linguistic correlations, allowing the model to predict without visual evidence. For questions containing semantic associations or stereotypical object-action relations, the model may assign high probability to a plausible but visually unsupported answer.

Salient-view probe. The salient-view probe keeps only a small number of locally relevant clips. Given a video v, we uniformly divide it into N non-overlapping temporal clips $\{c_i\}_{i=1}^{N}$, where each clip contains consecutive frames rather than an isolated frame. We then compute a query-conditioned relevance score a_i for each clip and select the top-K clips:

$$\mathcal{I}_{\text{sal}} = \text{TopK}_{i\in\{1,\dots,N\}}(a_i), \qquad v^s = \{c_i \mid i \in \mathcal{I}_{\text{sal}}\}, \tag{2}$$

where $\mathcal{I}_{\text{sal}}$ denotes the indices of the selected clips and v^s denotes the resulting salient-view input.

In our implementation, the clip-level score is computed from the model's attention response from the query tokens to the visual tokens of each clip:

$$a_i = \frac{1}{|\mathcal{Q}|\,|\mathcal{T}_i|} \sum_{q\in\mathcal{Q}} \sum_{t\in\mathcal{T}_i} A_{q\to t}, \tag{3}$$

where Q denotes the set of query tokens, T_i denotes the visual tokens belonging to clip c_i, and $A_{q\to t}$ denotes the attention response from query token q to visual token t. We use this score to select candidate clips for the salient-view probe.

The salient view is not intended to retrieve the complete evidence. Instead, it approximates the model's most immediately attended local context. Such local clips may contain objects, actions, or scenes that are highly related to the question, but they may still miss the temporal context required for the correct answer. Therefore, this probe exposes visually prominent but incomplete evidence.

To compare the full-video path with the restricted-view probes, we extract hidden states at the same textual decision position. For each view, we run a forward pass with the same textual query x and extract the hidden state of the shared decision token z before answer generation:

$$\mathbf{h}_l^m = H_l(v_m, x)[z], \qquad m \in \{f, q, s\}, \tag{4}$$

where $H_l(\cdot)$ denotes the hidden states at the l-th LLM decoder layer, and z is the last textual token before answer generation. We use the same prompt template across all views and extract $\mathbf{h}_l^f$, $\mathbf{h}_l^q$, and $\mathbf{h}_l^s$ at this fixed decision position, ensuring that only the visual observations differ.

4.3 Residual Contrastive Steering

The counterfactual probes above expose alternative trajectories formed under incomplete observations. CLEAR uses these trajectories as contrastive anchors to correct the full-video reasoning process. The key intuition is that the difference between the full-video trajectory and a shortcut-probe trajectory defines a direction from restricted-view reasoning toward complete-observation reasoning.

Shortcut residuals. Given the full-video hidden state $\mathbf{h}_l^f$ and a shortcut-probe hidden state $\mathbf{h}_l^m$ at layer l, where $m \in \{q, s\}$, we define:

$$\mathbf{r}_l^m = \mathbf{h}_l^f - \mathbf{h}_l^m. \tag{5}$$

This residual is not assumed to be a disentangled shortcut component. Instead, it is a directional contrast between incomplete-observation reasoning and full-video reasoning. In other words, it reflects the representational contrast between the restricted-view trajectory and the full-video trajectory.

For the two counterfactual probes, $\mathbf{r}_l^q$ provides a correction direction against spurious linguistic correlations, while $\mathbf{r}_l^s$ provides a correction direction against salient-yet-misleading visual observations.

Shortcut-repulsive update. CLEAR updates the full-video hidden state by injecting residual corrections from the activated shortcut probes:

$$\tilde{\mathbf{h}}_l^f = \mathbf{h}_l^f + \alpha_l^q \mathbf{r}_l^q + \alpha_l^s \mathbf{r}_l^s, \tag{6}$$

where the effective intervention strength for each probe is defined as

$$\alpha_l^m = \alpha G_l^m, \qquad G_l^m = \mathbb{I}(d_l^m > \tau), \qquad m \in \{q, s\}. \tag{7}$$

Here, α is a global steering coefficient, τ is a fixed discrepancy threshold, and G_l^m is a binary gate indicating whether the shortcut-probe trajectory sufficiently deviates from the full-video trajectory at layer l. Thus, a residual is injected only when the corresponding shortcut probe is activated.

Since each residual is defined as $\mathbf{r}_l^m = \mathbf{h}_l^f - \mathbf{h}_l^m$, it points from the restricted-view shortcut anchor to the full-video state. Adding activated residuals therefore moves the full-video state away from the shortcut anchors. The remaining decoder layers then continue computation from the corrected state.

Where and when to intervene. Although residual directions can be constructed at every layer, not all layers are equally suitable for correction.

We empirically compare intervention at different decoder blocks and find that late-layer intervention is the most effective, with the last block, layers 28–35, giving the best performance, as shown in Figure 4. This is consistent with the intuition that answer-oriented representations are formed near the decoding stage.

Within the selected layers, we further determine when to apply intervention based on the discrepancy between the full-video trajectory and the shortcut-probe

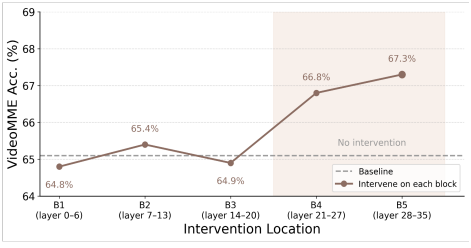


Figure 4: **Effect of intervention location.** Late-layer intervention performs best, with the last block achieving the highest accuracy.

Table 2: **General comparison on long-video understanding benchmarks.** CLEAR is evaluated on three open-source MLLMs, with closed-source and public models included for reference.

Model	Size	Frames	MLVU	Video-MME		LongVB	LVBench
				Overall	Long		
Closed-source MLLMs							
GPT-4o [28]	–	–	54.9	65.3	59.0	**66.7**	48.9
Gemini-1.5-Pro [35]	–	–	53.8	**67.4**	**61.8**	64.0	33.1
Open-source video MLLMs							
LongVU [31]	7B	1 fps†	63.7	60.9	54.3	59.5	–
NVILA [21]	8B	8	57.7	64.2	54.8	57.7	–
VideoLLaMA3 [48]	7B	1 fps†	63.0	66.2	–	59.8	45.3
mPLUG-Owl3 [46]	7B	256	62.1	64.3	53.0	58.7	32.6
ViLAMP [8]	7B	1 fps†	62.3	67.3	59.3	62.0	30.8
VideoChatFlash [17]	7B	16	64.7	65.3	55.4	64.7	48.2
LLaVA-Video [51]	7B	64	52.1	59.0	47.2	58.6	43.1
+ CLEAR	7B	64	55.3 ↑3.2	61.5 ↑2.5	49.8 ↑2.6	61.0 ↑2.4	49.0 ↑5.9
InternVL3 [55]	8B	64	57.8	63.1	52.7	59.4	43.4
+ CLEAR	8B	64	62.4 ↑4.6	66.7 ↑3.6	56.9 ↑4.2	62.3 ↑2.9	49.7 ↑6.3
Qwen3-VL [2]	8B	64	59.0	65.1	55.5	64.3	46.0
+ CLEAR	8B	64	**66.8** ↑7.8	**70.9** ↑5.8	**62.7** ↑7.2	**68.1** ↑3.8	**52.4** ↑6.4

† 1 fps: sampling one frame per second over the full video.

trajectory:

$$d_l^m = 1 - \cos(\mathbf{h}_l^f, \mathbf{h}_l^m). \tag{8}$$

A larger d_l^m indicates that the shortcut probe induces a more distinct reasoning trajectory. We therefore activate intervention only when $d_l^m > \tau$, using a fixed threshold $\tau = 0.15$ in all experiments.

5 Experiments

5.1 Experimental Setup

Models and Benchmarks. We evaluate CLEAR on representative open-source long-video MLLMs, including LLaVA-Video-7B, InternVL3-8B, and Qwen3-VL-8B. Closed-source models, including GPT-4o and Gemini-1.5-Pro, are reported only as reference systems since their hidden states are inaccessible. Experiments are conducted on four public long-video benchmarks, Video-MME, LongVideoBench, LVBench, and MLVU, covering general video QA, long-context temporal reasoning, and complex event-level understanding. We further evaluate on S-MME, our diagnostic benchmark for measuring whether models collapse to shortcut-consistent alternatives.

5.2 Main Results on Public Long-video Benchmarks

Compared to open-source baselines. As shown in Table 2, CLEAR consistently improves open-source long-video MLLMs under the same final-frame budget and decoding setting. The gains are most pronounced on reasoning-intensive benchmarks: CLEAR improves Qwen3-VL-8B by +7.8 on MLVU and +7.2 on Video-MME-Long. For InternVL3-8B, CLEAR improves Video-MME Overall by +3.6, Video-MME Long by +4.2, and LVBench by +6.3. On LLaVA-Video-7B, CLEAR also achieves a clear +5.9 improvement on LVBench.

These results show that CLEAR improves general long-video understanding without increasing the final-frame budget or changing the backbone. The stronger gains on MLVU and Video-MME-Long suggest that standard long-video benchmarks also contain evidence-use failures, where relevant evidence is present but models still rely on shortcut cues. Closed-source models are included as references, while CLEAR is applied to open-source models with accessible hidden states.

Table 3: **Diagnostic comparison on S-MME.** ShortErr denotes the percentage of predictions that incorrectly choose the annotated shortcut-consistent distractor.

Model	Size	Frames	Lang.-Prior		Part.-Evid.		Overall	
			Acc.↑	ShortErr.↓	Acc.↑	ShortErr.↓	Acc.↑	ShortErr.↓
Human Reference								
Human	–	–	92.4	2.6	90.1	3.8	91.2	3.2
Closed-source MLLMs								
GPT-4o	–	–	47.2	18.6	43.8	22.4	45.5	20.5
Gemini-1.5-Pro	–	–	54.1	14.2	50.3	17.9	52.2	16.1
Open-source video MLLMs								
LongVU	7B	1 fps†	15.6	51.8	12.4	56.7	14.0	54.3
NVILA	8B	8	17.9	48.6	14.8	53.1	16.4	50.9
VideoLLaMA3	7B	1 fps†	18.6	47.5	15.3	52.4	17.0	49.9
mPLUG-Owl3	7B	256	19.1	46.8	16.0	51.6	17.6	49.2
ViLAMP	7B	1 fps†	18.3	47.9	15.7	52.0	17.0	50.0
VideoChatFlash	7B	16	16.8	50.2	13.9	54.6	15.4	52.4
LLaVA-Video	7B	64	11.1	56.8	7.9	61.4	9.5	59.1
+ CLEAR	7B	64	14.8 ↑3.7	51.7 ↓5.1	10.6 ↑2.7	56.2 ↓5.2	12.7 ↑3.2	53.9 ↓5.2
InternVL3	8B	64	19.8	44.3	16.5	48.0	18.2	46.2
+ CLEAR	8B	64	24.6 ↑4.8	38.5 ↓5.8	18.9 ↑2.4	42.4 ↓5.6	21.7 ↑3.5	40.5 ↓5.7
Qwen3-VL	8B	64	22.2	41.7	18.8	46.2	20.5	44.0
+ CLEAR	8B	64	**28.4** ↑6.2	**34.1** ↓7.6	**24.5** ↑5.7	**39.6** ↓6.6	**26.5** ↑6.0	**36.9** ↓7.1

† 1 fps: sampling one frame per second over the full video.

Table 4: **Effect of shortcut probes on S-MME.** Q-only and S-only apply residual correction from the language-prior and salient-view probes, respectively. CLEAR combines both probes for complementary correction.

Variant	Q	S	Branch	Linguistic Bias		Visual Bias		Overall
				Acc.↑	ShortErr.↓	Acc.↑	ShortErr.↓	Acc.↑
Baseline	–	–	–	22.2	41.7	18.8	46.2	20.5
Q-only	✓	–	Single	26.1	36.8	20.4	43.5	23.3
S-only	–	✓	Single	24.5	38.7	22.9	41.0	23.7
CLEAR	✓	✓	Dual	**28.4**	**34.1**	**24.5**	**39.6**	**26.5**

5.3 Diagnostic Evaluation on S-MME

Table 3 shows that S-MME exposes substantial evidence-use failures in current long-video MLLMs. Human annotators reach 91.2% accuracy with only 3.2% ShortErr, while the strongest closed-source model achieves 52.2% accuracy and still produces 16.1% shortcut-consistent errors. Open-source models exhibit more severe shortcut collapse, with the best baseline Qwen3-VL reaching only 20.5% accuracy and 44.0% ShortErr. CLEAR improves both accuracy and shortcut robustness. On Qwen3-VL, it raises overall accuracy from 20.5% to 26.5% and reduces ShortErr from 44.0% to 36.9%. The reduction appears in both conflict types, suggesting that the two counterfactual probes address complementary shortcut sources.

5.4 Is More Context Enough?

A natural way to improve long-video understanding is to sample more frames. We compare the baseline and CLEAR under different frame budgets to test whether context scaling alone resolves shortcut-biased errors. On S-MME, we report both accuracy and ShortErr; on Video-MME, we report accuracy as a representative public benchmark. Figure 5 shows that more frames improve baseline accuracy, confirming the importance of evidence access. However, ShortErr decreases more slowly, indicating that context scaling alone does not fully resolve biased evidence use. Across all frame

Table 5: **Efficiency and intervention analysis.** (a) On Video-MME, CLEAR outperforms simple frame scaling under comparable inference budgets. (b) On S-MME, moderate residual strength yields the best accuracy–shortcut suppression trade-off.

(a) Efficiency on Video-MME

Method	Final F.	Aux. Probe	Time	Acc.↑
Qwen3-VL	32	–	0.62×	62.8
Qwen3-VL	64	–	1.00×	65.1
Qwen3-VL	128	–	1.86×	67.0
Qwen3-VL + CLEAR	32	Q+S	1.34×	68.4 ↑5.6
Qwen3-VL + CLEAR	64	Q+S	1.78×	**70.9** ↑5.8
Qwen3-VL + CLEAR	128	Q+S	2.72×	71.5 ↑4.5

(b) Intervention strength on S-MME

Method	α	Acc.↑	ShortErr.↓
Qwen3-VL	0.00	20.5	44.0
Qwen3-VL + CLEAR	0.05	23.4 ↑2.9	40.6 ↓3.4
	0.10	25.1 ↑4.6	38.5 ↓5.5
	0.15	**26.5** ↑6.0	36.9 ↓7.1
	0.25	26.0 ↑5.5	36.5 ↓7.5
	0.35	24.8 ↑4.3	**36.1** ↓7.9

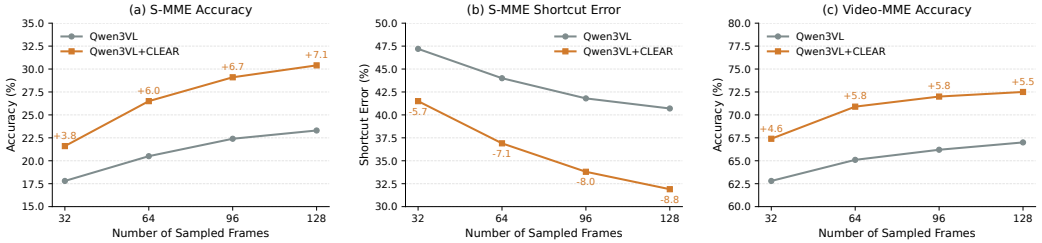


Figure 5: **Context scaling analysis.** CLEAR consistently improves accuracy and reduces ShortErr under the same frame budget.

budgets, CLEAR achieves higher accuracy and lower ShortErr, and its gains on Video-MME suggest that shortcut correction generalizes beyond S-MME.

5.5 Do Shortcut Probes Align with Conflict Types?

CLEAR uses two restricted views to expose different bias sources: the question-only probe targets spurious linguistic correlations, while the salient-view probe targets salient-yet-misleading visual observations. We compare single-branch and dual-branch variants on S-MME to examine whether these probes play source-specific roles. Table 4 shows that Q-only mainly improves cases dominated by spurious linguistic correlations, indicating its ability to capture question-option priors. S-only is more effective on cases involving salient-yet-misleading visual observations, suggesting that locally plausible visual cues cannot be diagnosed from language alone. Dual-branch variants further improve overall robustness, showing that the two restricted views provide complementary correction signals.

5.6 Efficiency and Intervention Analysis

Although CLEAR introduces auxiliary shortcut probes, its goal is not to simply increase inference compute. We compare it with a direct compute-scaling baseline that uses more full-video frames. Table 5(a) shows that increasing the full-video context brings limited gains with noticeably higher cost, while CLEAR achieves larger accuracy improvements under comparable inference budgets. This suggests that the benefit mainly comes from shortcut diagnosis and residual corrections.

We further study the strength of residual correction. Table 5(b) shows that a moderate residual scale gives the best accuracy–robustness trade-off. When α becomes too large, ShortErr continues to decrease slightly, but accuracy starts to drop, indicating over-correction. Overall, CLEAR remains training-free: the backbone is frozen, and the probes only provide inference-time correction signals.

6 Conclusion

We studied shortcut-driven failures in long-video MLLMs, where models may rely on spurious linguistic correlations or salient yet misleading visual observations despite the presence of decisive evidence. To diagnose this behavior, we introduced S-MME, a benchmark that tests whether models follow complete video evidence or collapse to shortcut-consistent alternatives. We further proposed

CLEAR, a training-free inference-time framework that uses counterfactual views to expose shortcut-induced trajectories and applies late-layer residual steering to the full-video path. Experiments on public long-video benchmarks and S-MME show that CLEAR consistently improves accuracy and reduces shortcut-consistent errors across different MLLMs.

Acknowledgments and Disclosures

Jiawei Ma acknowledges the support of the National Natural Science Foundation of China (NSFC) under Grant No. 62172439 and the Hong Kong Research Grants Council (RGC) under Grant No. TRS T43-101/26-N. This work was also supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 62422606 and 62441615, the Hong Kong Research Grants Council (RGC) General Research Fund (GRF) under Grant No. 17213925, and the Shenzhen Loop Area Institute Fund under Grant No. FPF10120250006. Hao Zhang acknowledges the support of the National Natural Science Foundation of China (NSFC) under Grant No. 62301384 and the Excellent Young Scientists Fund (Overseas).

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980, 2018.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[3] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32, 2019.

[4] Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20237–20246, 2025.

[5] Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Tuomas Rintamaki, et al. Eagle 2.5: Boosting long-context post-training for frontier vision-language models. *arXiv preprint arXiv:2504.15271*, 2025.

[6] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024.

[7] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10800–10809, 2020.

[8] Chuanqi Cheng, Jian Guan, Wei Wu, and Rui Yan. Scaling video-language models to 10k frames via hierarchical differential distillation. *arXiv preprint arXiv:2504.02438*, 2025.

[9] Dingxin Cheng, Mingda Li, Jingyu Liu, Yongxin Guo, Bin Jiang, Qingbin Liu, Xi Chen, and Bo Zhao. Enhancing long video understanding via hierarchical event-based memory. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2025.

[10] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language Processing (EMNLP-IJCNLP)*, pages 4069–4082, 2019.

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

[12] Yang Jin, Zhicheng Sun, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, et al. Video-LaVIT: Unified video-language pre-training with decoupled visual-motional tokenization. *arXiv:2402.03161*, 2024.

[13] Wenhao Li, Xianjing Meng, Qiangchang Wang, Zhongyi Han, Zhibin Wu, and Yilong Yin. Dvla-rl: Dual-level vision-language alignment with reinforcement learning gating for few-shot learning. In *ICLR*, 2026.

[14] Wenhao Li, Qiangchang Wang, Xianjing Meng, Zhibin Wu, and Yilong Yin. Vt-fsl: Bridging vision and text with llms for few-shot learning. In *NeurIPS*, volume 38, pages 113697–113727, 2025.

[15] Wenhao Li, Qiangchang Wang, Peng Zhao, and Yilong Yin. Knn transformer with pyramid prompts for few-shot learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1082–1091, 2024.

[16] Wenhao Li, Zhibin Wu, Chong Xiao, and Qiangchang Wang. Semmsa: Latent semantic-aided robust multimodal sentiment analysis with incomplete data. *arXiv preprint arXiv:2609.30238*, 2026.

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

[18] Xusheng Liang, Lihua Zhou, Nianxin Li, Miao Xu, Ziyang Song, Dong Yi, Jinlin Wu, Jiawei Ma, Hongbin Liu, Zhen Lei, et al. Multimodal causality-driven representation learning for generalizable medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13670–13679, 2026.

[19] Jingren Liu, Yun Wang, Long Zhang, Yiheng Wang, Shuning Xu, Ling Wang, Jiaqi Yan, Dell Zhang, and Xiangyu Chen. Towards training-free long video understanding: methods, benchmarks, and open challenges. *Vicinagearth*, 2(1):6, 2025.

[20] Xuyang Liu, Yiyu Wang, Junpeng Ma, and Linfeng Zhang. Video compression commander: Plug-and-play inference acceleration for video large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1910–1924, 2025.

[21] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4122–4134, 2025.

[22] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx MLLM: On-demand spatial-temporal understanding at arbitrary resolution. In *ICLR*, 2025.

[23] Jiawei Ma, Po-Yao Huang, Saining Xie, Shang-Wen Li, Luke Zettlemoyer, Shih-Fu Chang, Wen-Tau Yih, and Hu Xu. Mode: Clip data experts via clustering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26344–26353. IEEE, 2024.

[24] Jiawei Ma, Yulei Niu, Jincheng Xu, Shiyuan Huang, Guangxing Han, and Shih-Fu Chang. Digeo: Discriminative geometry-aware learning for generalized few-shot object detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3208–3218. IEEE, 2023.

[25] Wentao Ma, Weiming Ren, Yiming Jia, Zhuofeng Li, Ping Nie, Ge Zhang, and Wenhu Chen. Videoeval-pro: Robust and realistic long video understanding evaluation. *arXiv preprint arXiv:2505.14640*, 2025.

[26] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad 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 (Volume 1: Long Papers)*, pages 12585–12602, 2024.

[27] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710, 2021.

[28] OpenAI. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/, 2024. Accessed: 2024-05-20.

[29] Bin Pu, Xusheng Liang, Xinpeng Ding, Jinlin Wu, Zhen Lei, Shengli Li, Kenli Li, and Jiawei Ma. F^2-Assist: Multi-phase fetal growth forecast and report generation from ultrasound examination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 35341–35350, 2026.

[30] Saul Santos, António Farinhas, Daniel C. McNamee, and André Martins. ∞-Video: A training-free approach to long video understanding via continuous-time memory consolidation. In *International Conference on Machine Learning*, pages 52877–52893. PMLR, 2025.

[31] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal adaptive compression for long video-language understanding. *arXiv:2410.17434*, 2024.

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

[33] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024.

[34] Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29118–29128, 2025.

[35] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*, 2023.

[36] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv:2409.12191*, 2024.

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

[38] Yuxuan Wang, Yiqi Song, Cihang Xie, Yang Liu, and Zilong Zheng. Videollamb: Long streaming video understanding with recurrent memory bridges. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24170–24181, 2025.

[39] Zikang Wang, Boyu Chen, Zhengrong Yue, Yi Wang, Yu Qiao, Limin Wang, and Yali Wang. Videochat-a1: Thinking with long videos by chain-of-shot reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 10467–10475, 2026.

[40] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3272–3283, 2025.

[41] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

[42] Jinlin Wu, Xusheng Liang, Xuexue Bai, and Zhen Chen. Surgbox: Agent-driven operating room sandbox with surgery copilot. In *2024 IEEE International Conference on Big Data (BigData)*, pages 2041–2048. IEEE, 2024.

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

[44] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. PLLaVA: Parameter-free llava extension from images to videos for video dense captioning. *arXiv:2404.16994*, 2024.

[45] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. SlowFast-LLaVA: A strong training-free baseline for video large language models. *arXiv:2407.15841*, 2024.

[46] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024.

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

[48] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

[49] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv:2406.16852*, 2024.

[50] Yiming Zhang, Zhuokai Zhao, Zhaorun Chen, Zenghui Ding, Xianjun Yang, and Yining Sun. Beyond training: Dynamic token merging for zero-shot video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22046–22055, 2025.

[51] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[52] Yunzhu Zhang, Yu Lu, Tianyi Wang, Fengyun Rao, Yi Yang, and Linchao Zhu. Flexselect: Flexible token selection for efficient long video understanding. *arXiv preprint arXiv:2506.00993*, 2025.

[53] Zhuo Zhi, Qiangqiang Wu, Wenbo Li, Yinchuan Li, Kun Shao, Kaiwen Zhou, et al. Videoagent2: Enhancing the llm-based agent system for long-form video understanding by uncertainty-aware cot. *arXiv preprint arXiv:2504.04471*, 2025.

[54] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2(5):6, 2024.

[55] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[56] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. Apollo: An exploration of video understanding in large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18891–18901, 2025.

A Technical appendices and supplementary material

A.1 Limitation

CLEAR is training-free but not cost-free. Since it uses question-only and salient-view probes to estimate shortcut residuals, each sample requires additional auxiliary forward passes before final full-video decoding. This multi-perspective inference introduces extra latency and computation compared with a vanilla single-pass baseline, particularly for long videos and large MLLM backbones.

Our current focus is to validate the effectiveness of controlled shortcut probes for evidence-grounded reasoning. Improving the efficiency of CLEAR is an important future direction. Potential solutions include adaptive probe activation, lightweight probe approximations, representation caching, shared early-layer computation, and early exiting when the discrepancy between the full-video and shortcut-probe trajectories is small.

A.2 Implementation Details

All experiments are conducted with frozen open-source MLLMs, including LLaVA-Video-7B, InternVL3-8B, and Qwen3-VL-8B. Closed-source models are only reported as reference systems since their hidden states are inaccessible. Unless otherwise specified, we uniformly sample 64 frames from each video for the final full-video inference path. For CLEAR, we run two auxiliary probes before final decoding: a question-only probe and a salient-view probe. The question-only probe removes all visual inputs, while the salient-view probe selects the top-K query-relevant clips according to attention scores from query tokens to visual tokens; we use $K = 4$ by default. The probes are only used to compute shortcut residuals and never directly generate the final answer. For each view, hidden states are extracted at the same textual decision position.

Residual steering is applied to late decoder layers. For Qwen3-VL-8B, we use layers 28–35 by default, and for other backbones we use the corresponding late-layer block with the same relative depth. A probe-specific residual is activated only when $d_l^m = 1 - \cos(h_l^f, h_l^m) > \tau$, with $\tau = 0.15$ for all experiments. The default residual strength is $\alpha = 0.15$, selected from $\{0.05, 0.10, 0.15, 0.25, 0.35\}$ according to the accuracy–shortcut suppression trade-off.

All experiments are run on NVIDIA H100 GPUs with batch size one for long-video inference. Efficiency is reported as relative inference time normalized by the vanilla 64-frame baseline of the same backbone, including auxiliary probing, residual construction, and final full-video decoding.

Open-ended video QA. To examine whether CLEAR is limited to multiple-choice prediction, we further evaluate it on VideoEvalPro, an open-ended video question answering benchmark. As shown in Table 6, CLEAR consistently improves free-form generation across all three backbones, with gains of +1.2, +1.5, and +3.4 points on LLaVA-Video-7B, InternVL3-8B, and Qwen3-VL-8B, respectively. Since VideoEvalPro requires models to generate answers without candidate options, these improvements suggest that CLEAR does not merely calibrate multiple-choice scores, but also steers generation toward more evidence-consistent responses.

Table 6: **Open-ended evaluation on VideoEvalPro[25].** CLEAR improves free-form video question answering, suggesting that its gains are not limited to multiple-choice task.

Model	Frames	Method	VideoEvalPro
LLaVA-Video-7B	64	Baseline	22.8
LLaVA-Video-7B	64	+ CLEAR	24.0
InternVL3-8B	64	Baseline	23.8
InternVL3-8B	64	+ CLEAR	25.3
Qwen3-VL-8B	64	Baseline	25.2
Qwen3-VL-8B	64	+ CLEAR	28.6

Qualitative analysis. Figure 6 shows an open-ended QA example where CLEAR corrects a shortcut-biased response and produces an answer better aligned with the decisive temporal evidence.



What attack did Superman launch on SpongeBob first?

Qwen3VL-8B	***Qwen3VL-8B+CLEAR***
Superman charges at SpongeBob, attacks him, and traps him in **a tornado**.	Superman attempts to use **"Water Balls"** as a long-range projectile weapon. However, the attack is completely ineffective.

Figure 6: Qualitative effect of CLEAR on open-ended video QA.

A.3 Broader Impact

CLEAR operates as a training-free inference-time module that steers long-video MLLMs toward evidence-consistent predictions. It may improve reliability in long-form video understanding tasks, but it cannot prevent the base MLLMs from generating erroneous, biased, or harmful responses. Therefore, CLEAR should be viewed as a complementary robustness technique rather than a standalone safety mechanism, especially in high-stakes applications.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We carefully described our contributions in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed our limitations in appendix A.1

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper is not about theory.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided details about our methods and implementation in the main paper and appendix. We will also upload our main code as supplementary materials.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We plan to release the CLEAR scripts, S-MME annotations and evaluation scripts

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We present the experimental setup and details in the main paper

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: However, we provide our code, detail configuration and random seeds to facilitate the reproducibility of our results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We clarify our computation resources at implementation details in appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We carefully reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed societal impact in appendix A.3.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The data and models used in this paper have been extensively used in the multi-modal understanding community, and have undergone comprehensive safety risk assessments.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We provided appropriate citations in the reference section.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have included the code and benchmark.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not conduct human-subject experiments. Human involvement is limited to internal annotation and verification of video-QA samples; no personal data, behavioral measurements, or user studies are collected.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human-subject experiments. The human annotation process is used only for dataset verification and does not collect personal or sensitive information from annotators.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method developed in this study was derived through extensive hypothesis-experimentation cycles, without utilizing LLMs as any essential, novel, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.