SITE: SURFACING INTER-FRAME TEMPORAL EVIDENCE TO RECOVER VLMS' DORMANT TEMPORAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) typically answer temporally discriminative video questions using uniformly sampled frames. This interface is efficient, but presents the same frame-wise input regardless of the temporal relation required by the question. Even when the frames contain the relevant visual observations, these relations remain implicit. Answering correctly therefore requires the VLM to both organize these observations into usable temporal evidence and reason over that evidence. These two operations are coupled within a single prediction, so final-answer accuracy alone cannot distinguish failures in evidence formation from failures in answer reasoning. We investigate whether explicitly presenting question-relevant temporal evidence can correct errors in frozen-VLM under this interface. We introduce SITE, a training-free Route–Surface–Corroborate framework. SITE routes each question to the required type of inter-frame evidence, constructs corresponding evidence views through visual organization and task-specific instructions, and combines their predictions through cross-view voting. Across three benchmarks and three frozen backbones, SITE improves accuracy by up to 23.0 percentage points over the sampled-frames baseline and outperforms matched-backbone training-free baselines. These results show that explicitly organizing relevant temporal evidence and corroborating predictions across views can correct temporal QA errors without updating model parameters.

1 INTRODUCTION

Vision-Language Models (VLMs) commonly answer temporally discriminative video questions using a compact set of uniformly sampled frames (Li et al., 2026; Liu et al., 2024; Cai et al., 2024). These frames are ordinarily presented as the same chronological sequence regardless of the temporal relation queried. This generic presentation retains the sampled frames and their order, but leaves question-specific relations among them implicit.

Answering from this input couples two operations within a single prediction. The VLM must first identify the relevant sampled frames and organize the relations among them into *question-relevant temporal evidence*. It must then reason over this evidence to produce an answer. Standard evaluation scores only the final answer, so errors in evidence formation and answer reasoning are absorbed into a single accuracy measure. The resulting score does not distinguish a failure to form usable temporal evidence from a failure to reason over that evidence.

Existing approaches to temporal QA broadly fall into training-based and training-free methods. Training-based methods update model parameters to learn temporal representations or reasoning policies (Ren et al., 2024; Huang et al., 2024a; Qian et al., 2024; Huang et al., 2024b; Wang et al., 2024a; Feng et al., 2026). Training-free methods instead adapt frame selection, seek additional evidence, annotate the input, or introduce further inference procedures (Xu et al., 2026; Wang et al., 2024b; 2025; Shang et al., 2024). These methods demonstrate the value of stronger temporal reasoning and more targeted access to video evidence. We examine a complementary way to support reasoning by organizing the available sampled frames to make the temporal relations required by the question explicit.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Rather than leaving this operation implicit within answer inference, we treat evidence formation as an explicit stage between video sampling and answer generation. Given a question and the sampled frames, this stage constructs an *evidence view* that presents the inter-frame relation required for reasoning. This intervention tests whether explicit evidence presentation can help the frozen VLM correct errors made under the standard sampled frame input. Such improvements would indicate that organizing the available observations into question-relevant evidence can better elicit the model's existing temporal reasoning capabilities.

We introduce SITE (**S**urfacing **I**nter-frame **T**emporal **E**vidence) to instantiate this stage while keeping the VLM frozen. As shown in Fig. 1, it uses a training-free *Route–Surface–Corroborate* pipeline. **Route** infers from the question the form of temporal evidence required for the answer. **Surface** constructs one or more evidence views that present the required inter-frame relation explicitly. **Corroborate** obtains an independent answer from each view and resolves disagreements. Together, these stages adapt evidence presentation to each question while retaining a frozen VLM backbone.

Experiments across temporally discriminative video QA benchmarks and frozen VLM backbones show that SITE consistently improves standard uniform-frame inference and outperforms training-free methods using the same backbone. SITE also narrows performance gaps to post-trained models where the original backbones underperform. Experiments with post-trained models further show that SITE can complement temporal post-training without additional parameter updates. Further analysis shows that organizing temporal evidence around the question and combining predictions from different presentations both contribute to the gains. SITE also corrects a substantial fraction of errors made under the standard sampled frame input, even for the highest-performing backbone. These findings suggest that existing temporal reasoning capabilities can be better elicited through question-relevant evidence presentation without additional training.

Our contributions are:

- We investigate how explicitly presenting the temporal relations required by the question at the input interface can better elicit temporal reasoning from frozen VLMs.
- We introduce SITE, a Route–Surface–Corroborate framework that identifies the required evidence form, constructs the corresponding evidence views, and corroborates their answers without updating the VLM.
- We demonstrate consistent gains across benchmarks and frozen VLM backbones, outperforming matched-backbone training-free baselines while remaining competitive with post-trained models. SITE corrects a substantial fraction of baseline errors even for the highest-performing backbone.

2 Related Work

Temporally discriminative video QA. Temporally discriminative video QA involves questions that require reasoning about inter-frame relations and cannot be answered from static visual content alone. Controlled benchmarks such as TimeBlind (Li et al., 2026), DyBench (Du et al., 2026), VinoGround (Zhang et al., 2024), and VITATECS (Li et al., 2024b) use paired videos or descriptions to reduce static shortcuts and isolate temporal differences. Broader suites including TVBench (Cores et al., 2024), MVBench (Li et al., 2024a), MotionBench (Hong et al., 2025), TempCompass (Liu et al., 2024), TemporalBench (Cai et al., 2024), and TOMATO (Shangguan et al., 2025) cover complementary temporal skills and video settings. Together, these benchmarks show that current VLMs remain sensitive to how temporal relations are represented. They provide the temporally discriminative tasks on which we study evidence formation in frozen VLMs.

Learning temporal reasoning. Temporal reasoning can be learned directly in model parameters through supervised instruction tuning, temporal grounding objectives, or post-training. TimeChat (Ren et al., 2024) introduces timestamp-aware video instruction tuning for long-video understanding, while VTimeLLM (Huang et al., 2024a) and LITA (Huang et al., 2024b) train models to associate language with video moments and temporal locations. Momentor (Qian et al., 2024) uses fine-grained temporal supervision, and Grounded-VideoLLM (Wang et al., 2024a) strengthens the connection between generated language and localized video segments. These methods teach the model to represent time and grounding targets as part of its learned prediction process.

Recent post-training methods use reinforcement objectives and counterfactual supervision to improve spatiotemporal reasoning (Li et al., 2025; Feng et al., 2026; Du et al., 2026). The resulting capability is encoded in the updated model weights, which can benefit many downstream prompts once training is complete. This approach requires training data, parameter access, and a checkpoint produced for the chosen backbone. SITE addresses the same temporally discriminative questions under a different constraint: the VLM remains fixed, and evidence formation is changed only at inference time.

Training-free temporal reasoning. Training-free methods retain the model weights but change how video evidence is accessed, selected, or represented at inference time. Agentic methods such as the evidence-seeking agent that (Xu et al., 2026) developed for the TimeLogic Challenge (hereafter TimeLogic) and VideoAgent (Wang et al., 2024b) search the video iteratively and use intermediate model outputs to decide which moments to inspect next. VideoTree (Wang et al., 2025) builds a hierarchical representation of long videos; TraveLER (Shang et al., 2024) and MoReVQA (Min et al., 2024) divide video QA into modular perception and reasoning operations. These methods allocate inference-time computation according to the question instead of committing to one fixed set of sampled frames.

Another group changes the input presented to the VLM. Adaptive keyframe sampling (Tang et al., 2025), BOLT (Liu et al., 2025), and Frame-Voyager (Yu et al., 2025) select or query compact frame sets before answering. ViKey retains a visual prompting interface, adding frame indices and mapping question keywords to candidate keyframes (Lee et al., 2026). NeuS-QA takes a different route: it translates perceived events into symbolic sequences and evaluates temporal-logic expressions over that representation (Shah et al., 2026). More general program-mediated reasoning and visual prompting methods likewise structure intermediate visual evidence without updating the target model (Surís et al., 2023; Gupta & Kembhavi, 2023; Yang et al., 2023).

SITE also belongs to this training-free family. Unlike methods that acquire additional moments or convert video into symbolic programs, it operates on the sampled frames and changes how question-relevant inter-frame evidence is presented to a frozen VLM.

3 METHOD

3.1 OVERVIEW

Given a video V, a temporally discriminative question q, a set of answer choices $\mathcal{A}$, and a frozen VLM M, the standard inference interface uniformly samples an ordered sequence of T frames

$$\mathcal{F} = (f_{t_1}, \ldots, f_{t_T}), \qquad t_1 < \cdots < t_T, \tag{1}$$

where f_{t_i} denotes the frame sampled at time t_i. The model predicts

$$a_{\text{base}} = M(q, \mathcal{A}, \mathcal{F}), \qquad a_{\text{base}} \in \mathcal{A}, \tag{2}$$

where a_{base} is the baseline prediction obtained from the uniformly sampled frame sequence. The sampled frames are presented in chronological order, but this generic sequence does not explicitly represent the temporal relation required to answer q. The model must therefore identify the relevant observations, organize their inter-frame relations into question-relevant temporal evidence, and infer the answer within a single prediction.

To decouple evidence formation from answer reasoning, SITE makes evidence presentation an explicit, question-conditioned stage through a *Route–Surface–Corroborate* pipeline, as shown in Fig. 1. In the **Route** stage, a question-only router maps q to one of five inter-frame evidence types according to the temporal relation required by the question. In the **Surface** stage, the sampled frames are used to construct one or more VLM inputs for the selected evidence type. Each such input is an evidence view that organizes the frames, adds temporal annotations, or derives a motion cue to make the required inter-frame relation explicit. Finally, in the **Corroborate** stage, SITE combines evidence-view predictions with an additional independent prediction when the predictions across views disagree. The answer with the highest vote count is selected, and a disagreement verifier determines the final answer when the vote is tied.

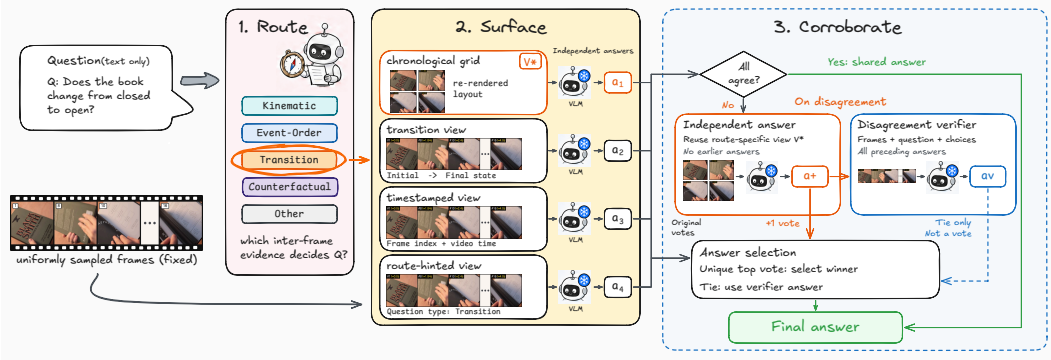

Figure 1: **Overview of SITE.** Given a question, SITE constructs complementary views for the selected evidence type. The *Transition* example shows a chronological grid, a transition view, a timestamped view and a route-hinted view. The grid arranges frames chronologically and the timestamped view adds frame indices and video times. The other two views modify only the prompt to request evidence of state change or provide the predicted route cue. Their labels describe prompt changes rather than frame annotations. The frozen VLM answers each view independently. Corroborate returns the shared answer if all predictions agree. Otherwise, it queries the highest priority view (V^*) again without access to earlier answers. The original predictions and this additional answer each contribute one vote. The answer with the most votes is selected. Ties are resolved by a verifier that examines the frames, question, choices and preceding predictions but does not vote.

Table 1: Evidence views constructed for each route. The highest-priority view (Route-Specific View) is listed first and highlighted in bold.

Route	Evidence views
Kinematic	**Optical-flow view**, Route-hinted view
Event-Order	**Question-selected ordering view**, Timestamped view, Route-hinted view
Transition	**Chronological grid**, Transition view, Timestamped view, Route-hinted view
Counterfactual	**Counterfactual view**, Timestamped view, Route-hinted view
Other	**Timestamped view**, Route-hinted view

3.2 Evidence Routing

Across temporally discriminative video QA benchmarks, questions repeatedly involve a compact set of inter-frame relations: object motion, event order, state or scene transition, and counterfactual outcomes (Li et al., 2026; Du et al., 2026; Liu et al., 2024; Li et al., 2024a; Hong et al., 2025). As shown in Table 1, we organize these relations into four specialized routes, with *Other* for questions outside these categories.

The frozen VLM backbone also serves as a text-only router. It receives the question and brief definitions of the five routes, then selects the route that describes the temporal relation required to answer the question. The model returns only a route label without receiving video frames, answer options, or benchmark identity. The selected route determines which evidence views Surface constructs.

3.3 Evidence Surfacing

Surface stage receives the route selected in Sec. 3.2 and constructs a set of *evidence views* from the sampled frames. Each view presents the inter-frame relation selected by the route through a particular form of visual evidence and a corresponding instruction. Depending on the route, Surface constructs two to four evidence views per question.

Two evidence views recur across routes. The route-hinted view preserves the sampled frames and changes only the prompt by stating the question type predicted by the router. The timestamped view adds the frame index and video time directly to each sampled frame. The frames remain in chronological order.

The remaining views are specialized for the selected route. The question-selected ordering view uses timestamped visual input and an ordering instruction that asks the VLM to locate the queried events and compare their temporal positions. The transition view keeps the visual input unchanged and modifies only the prompt. It asks the VLM to identify the initial state, final state, and visual evidence of change. The chronological grid changes the visual layout by arranging the sampled frames in a time-ordered grid. The counterfactual view also keeps the visual input unchanged and modifies only the prompt. It asks the VLM to ground the observed events before reasoning about the stated intervention. The Other route uses only the route-hinted view and timestamped view.

The optical-flow view used by Kinematic adds motion-annotated frames and a displacement summary produced by optical-flow analysis of the full video. Table 1 summarizes the views constructed for every route. For each route, we designate the highest-priority view in the table as the *route-specific evidence view*. The frozen VLM evaluates each evidence view together with the original question and answer choices. Corroborate then receives the resulting predictions. Further details on evidence view construction are provided in Appendix A.2.

3.4 Cross-View Corroboration

The Surface stage constructs multiple evidence views for the selected route. Each view presents the relevant inter-frame relation through a different visual organization or instruction. The frozen VLM answers the question independently from each view, but these predictions may disagree. Corroborate combines the resulting predictions rather than relying on a single evidence presentation.

If all views agree, Corroborate returns their shared answer. Otherwise, the frozen VLM produces an additional prediction using the route-specific evidence view defined in Sec. 3.3, without access to the earlier answers. The original predictions and the additional prediction each contribute one vote. An answer with a unique highest vote count is returned directly.

Only when multiple answers remain tied for the highest vote count does Corroborate invoke the disagreement verifier. The verifier receives the sampled frames, original evidence views, question and answer choices, and all preceding predictions. It evaluates the competing answers against the visual evidence and returns the final answer. Its prediction is not included in the vote count. Further details on cross-view corroboration are provided in Appendix A.3.

4 Experiments

4.1 Setup

Benchmarks. We evaluate SITE on **TimeBlind** (Li et al., 2026), **DyBench** (Du et al., 2026), and **VTL-Core**. **TimeBlind** contains 600 groups (2,400 records). Each group combines two complementary questions with two videos that share static content but differ in the target temporal factor, yielding four independently answered records. We report four metrics of increasing strictness: item-level accuracy (Acc); question accuracy (Q-Acc), which requires a question to be answered correctly on both videos of its pair; video accuracy (V-Acc), which requires both questions to be answered correctly on the same video; and instance accuracy (I-Acc), which requires all four records in a group to be correct. Q-Acc and V-Acc therefore measure the two directions of paired temporal discrimination separately, while I-Acc requires both at once. **DyBench** contains 1,507 pairs (3,014 instances) spanning reversible dynamics, moving direction, and event sequence. An original video and its controlled counterpart share the scene, question, and answer choices, while a spatiotemporal transformation changes the relevant dynamics and correct answer. We report item accuracy (Acc) and pair accuracy (P-Acc); the latter counts a pair as correct only when both constituent items are answered correctly. We additionally construct **VTL-Core**, a 470-question evaluation set containing 60 counterfactual inference, 200 scene transition, and 150 moving direction questions from MVBench (Li et al., 2024a), together with 60 Action Order questions from MotionBench (Hong et al., 2025). VTL-Core is evaluated using item accuracy (Acc).

Backbones. We evaluate SITE on three frozen VLM backbones with different multimodal reasoning capabilities: **Qwen3.6-27B** (Qwen Team, 2026), **MiniMax-M3** (MiniMax, 2026), and **GPT-5.6-Terra** (OpenAI, 2026). These backbones span distinct model families and differ in their multimodal

reasoning capabilities, allowing us to test whether SITE generalizes beyond a particular architecture or capability regime. We apply the same routing, evidence surfacing, and corroboration protocol to all three backbones.

Baselines. We compare SITE against three representative baseline families. For controlled same-backbone comparisons, we include: (i) **Base**, the frozen VLM evaluated in its native inference setting without SITE; (ii) the agent developed by (Xu et al., 2026) for the TimeLogic challenge (denoted **TimeLogic**), which iteratively acquires timestamped observations through sampling tools at multiple temporal scales without additional training; and (iii) **ViKey** (Lee et al., 2026), which overlays ordinal indices on the sampled frames and uses Keyword–Frame Mapping to associate query concepts with relevant frame indices, thereby providing explicit temporal anchors. We also include four post-trained video reasoning models for additional comparisons: **Video-R1** (Feng et al., 2026), which uses T-GRPO to contrast ordered and shuffled frame sequences, rewarding policies that rely on temporal order; **CRPO** (Du et al., 2026), which post-trains LVLMs with dual-branch counterfactual reinforcement to enhance spatiotemporal sensitivity; **VideoTemp-o3** (Liu et al., 2026), which combines temporal grounding with agentic video reasoning; and **TPRU** (Gao et al., 2026), which targets temporal and procedural understanding by constructing training signals from temporal reordering, next-frame prediction, and previous-frame review to post-train VLM with Group Relative Policy Optimization (GRPO) (Shao et al., 2024). These baselines cover agentic evidence seeking, frame annotation, and reinforcement-learned reasoning, enabling us to compare SITE against both input-side and model-side approaches.

Inference settings. All VLM backbones remain frozen and use their native reasoning modes. Base and SITE receive 16 uniformly sampled frames. Base and SITE calls use temperature 0 except for the additional independent prediction in Corroborate, which is sampled at temperature 0.7 without access to earlier answers. These settings remain fixed across benchmarks for each backbone. All other methods retain their original inference procedures.

4.2 Main results

Overall gains. Table 2 presents the main results. SITE consistently improves over Base across all three benchmarks and VLM backbones. For example, with MiniMax-M3, it raises TimeBlind I-Acc from 36.0% to 51.3% and VTL-Core accuracy from 54.5% to 73.2%. The gains persist with Qwen3.6-27B and GPT-5.6-Terra, demonstrating SITE's effectiveness across model families and suggesting that explicit temporal evidence organization provides a broadly applicable strategy for improving video understanding.

Comparison with Training-free Methods. Among methods using the same frozen backbone, SITE achieves the best result on every reported metric. Its advantage over TimeLogic shows that structuring evidence according to the question can be more effective than searching for additional frames alone. This also suggests that uniformly sampled observations contain underexploited information, and explicitly surfacing question-relevant temporal evidence allows VLMs to make better use of these observations. ViKey, by contrast, falls short of SITE and even underperforms Base in some settings, suggesting that temporal anchoring alone may not suffice to resolve how events unfold across frames.

Comparison with Post-trained Models. We compare with post-trained models to examine whether SITE can narrow performance gaps where VLM backbones initially underperform. On VTL-Core, the post-trained models achieve accuracies ranging from 66.2% to 71.1%. MiniMax-M3 and Qwen3.6-27B initially fall below this range at 54.5% and 47.0%. Applying SITE raises their accuracies to 73.2% and 70.0%. MiniMax-M3 exceeds the strongest post-trained result and Qwen3.6-27B approaches it. These results suggest that SITE can help the evaluated backbones achieve competitive accuracy without additional training.

To examine whether SITE remains effective on models already post-trained for temporal understanding, we evaluate SITE with TPRU-7B and TPRU-32B because their image sequence interface directly accommodates SITE's evidence views. Video-R1 and CRPO use native video inputs in our evaluation, while VideoTemp-o3 also uses video inspection tools. Directly replacing their video inputs or tool interactions with SITE's image sequences would introduce a mismatch between the formats used during temporal post-training and inference. As shown in Table 2, SITE improves

Table 2: **Main results on TimeBlind, VTL-Core, and DyBench.** All values are accuracies (%; higher is better). Methods within each backbone block use the same frozen backbone and evaluation items. The best and second-best results within each block are bolded and underlined, respectively.

Backbone	Method	TimeBlind				VTL-Core	DyBench	
		Acc	Q-Acc	V-Acc	I-Acc	Acc	Acc	P-Acc
MiniMax-M3	Base	73.0	53.3	57.6	36.0	54.5	67.3	47.1
	TimeLogic	<u>77.0</u>	<u>60.4</u>	<u>67.3</u>	<u>48.7</u>	<u>69.8</u>	<u>69.0</u>	<u>50.2</u>
	ViKey	72.0	51.2	55.8	35.0	64.9	59.5	39.0
	SITE (Ours)	**78.6**	**62.1**	**68.7**	**51.3**	**73.2**	**73.7**	**55.1**
Qwen3.6-27B	Base	64.3	47.1	53.8	37.8	47.0	69.9	52.0
	TimeLogic	<u>72.7</u>	<u>54.7</u>	<u>59.9</u>	<u>42.2</u>	<u>63.6</u>	<u>74.9</u>	<u>59.4</u>
	ViKey	69.0	49.1	56.9	39.5	58.9	73.1	57.9
	SITE (Ours)	**77.6**	**61.1**	**68.1**	**51.2**	**70.0**	**76.9**	**61.7**
GPT-5.6-Terra	Base	76.0	56.8	61.0	40.3	73.6	82.6	69.1
	TimeLogic	<u>81.2</u>	<u>63.8</u>	<u>70.8</u>	<u>55.7</u>	<u>78.1</u>	<u>83.9</u>	<u>71.7</u>
	ViKey	72.1	51.0	57.4	36.8	69.1	71.0	52.1
	SITE (Ours)	**81.7**	**67.6**	**72.3**	**56.5**	**82.6**	**87.9**	**79.0**
Post-trained Models	Video-R1-7B	59.1	21.8	36.3	13.3	67.9	60.5	36.8
	CRPO-8B	71.7	45.1	54.4	36.0	66.2	70.8	55.5
	VideoTemp-o3	59.2	32.8	39.7	15.0	71.1	63.1	41.0
	TPRU-7B	61.9	28.2	40.8	19.2	66.4	62.5	40.3
	TPRU-32B	63.3	30.5	43.9	21.2	66.4	63.5	41.4
Post-trained Models + SITE (Ours)	**TPRU-7B+SITE**	62.7	31.4	42.8	19.5	75.7	62.3	40.2
	TPRU-32B+SITE	65.5	37.3	47.3	25.5	73.8	66.4	44.3

TPRU-7B on TimeBlind and VTL-Core but slightly reduces its performance on DyBench. The slight decline on DyBench reflect limitations in the model's perceptual and temporal reasoning capabilities, whereas SITE yields larger gains with TPRU-32B across all three benchmarks. These results suggest that making question-relevant temporal evidence explicit remains beneficial even for models specifically trained for temporal understanding.

4.3 Ablation and Analysis

Table 3: Stage-wise ablation of SITE across backbones on TimeBlind, VTL-Core, and DyBench.

Inference condition	TimeBlind				VTL-Core	DyBench	
	Acc	Q-Acc	V-Acc	I-Acc	Acc	Acc	P-Acc
MiniMax-M3							
Sampled Frames	73.0	53.3	57.6	36.0	54.5	67.3	47.1
+ Route-Hinted View	73.4	55.0	60.6	40.2	55.5	67.8	47.9
+ Route-Specific View	75.8	59.0	64.9	48.2	68.7	68.3	49.2
+ Corroboration	**78.6**	**62.1**	**68.7**	**51.3**	**73.2**	**73.7**	**55.1**
Qwen3.6-27B							
Sampled Frames	64.3	47.1	53.8	37.8	47.0	69.9	52.0
+ Route-Hinted View	65.7	48.5	55.7	39.0	48.1	70.2	54.8
+ Route-Specific View	71.8	54.4	61.3	44.8	53.4	76.3	60.7
+ Corroboration	**77.6**	**61.1**	**68.1**	**51.2**	**70.0**	**76.9**	**61.7**
GPT-5.6-Terra							
Sampled Frames	76.0	56.8	61.0	40.3	73.6	82.6	69.1
+ Route-Hinted View	77.7	60.1	64.9	46.8	74.0	82.7	70.2
+ Route-Specific View	79.5	64.2	68.6	52.0	76.4	86.4	76.4
+ Corroboration	**81.7**	**67.6**	**72.3**	**56.5**	**82.6**	**87.9**	**79.0**

Component Contributions. To determine how each component contributes to SITE's performance, we progressively add the Route-Hinted View, Route-Specific View, and Corroboration to Sampled frames. The Route-Hinted View retains the same frames and changes only the prompt by stating the question type predicted by the router. Route-Specific View uses the highest-priority evidence view for the selected route to answer the question with a single model call, without corroboration.

Table 3 shows a consistent progression through the three stages. The Route-Hinted View provides an initial benefit, while the Route-Specific View yields larger subsequent gains. These results suggest

that explicitly surfacing task-relevant visual evidence is more effective than merely indicating the question type in the prompt. Finally, Corroboration consistently improves every reported metric across all three backbones. These gains highlight its role in reconciling predictions across multiple renderings of the selected evidence type, and producing more robust predictions.

SITE Gains Across Evidence Routes. Table 4 compares SITE's gains across evidence routes. The four specialized routes cover 84.98% of questions, while Other accounts for 15.02%. Full SITE improves over Sampled Frames for every route, with gains ranging from 5.77% to 10.50%. Kinematic achieves the largest gain of 10.50%, while the other four routes without derived visual cues improve by 5.77% to 6.82%. The 5.77% gain on Other suggests that SITE remains beneficial when the router does not select a specialized evidence relation. These gains suggest that SITE improves reasoning across diverse temporal relations, not only Kinematic questions supported by optical-flow visualizations.

Table 4: MiniMax-M3 accuracy (Acc, %) grouped by evidence route across the three benchmarks. Share denotes each route's percentage of all questions pooled across these benchmarks. Arrows indicate gains of full SITE over *Sampled Frames*. Kinematic additionally uses optical-flow visualizations; the remaining routes use only the sampled frames.

Route	Share (%)	Sampled Frames	Full SITE
Transition	36.64	80.10	86.92 ↑ 6.82
Event-Order	18.44	61.75	67.74 ↑ 5.99
Counterfactual	15.01	55.83	62.63 ↑ 6.80
Other	15.02	69.68	75.45 ↑ 5.77
Kinematic	14.89	60.62	71.12 ↑ 10.50

Error Recovery Across Backbones. We examine how SITE changes answer correctness relative to Sampled Frames across backbones in Fig. 2. Each sample falls into one of four categories: correct under both conditions (C→C), corrected by SITE (W→C), changed from correct to incorrect by SITE (C→W), or incorrect under both conditions (W→W). Although SITE changes some correct baseline answers to incorrect ones (C→W), it corrects more errors (W→C) than it introduces across all evaluated backbones and benchmarks. The recovery curves further show that SITE corrects a substantial fraction of baseline errors across all evaluated settings. GPT-5.6-Terra has the highest Sampled Frames accuracy on all three benchmarks, yet SITE still corrects 48.3% of its baseline errors on TimeBlind, 56.5% on VTL-Core and 38.3% on DyBench. These results show that stronger backbone performance does not eliminate errors that can be corrected through evidence surfacing. Explicitly organizing temporal evidence remains beneficial even for the strongest backbone evaluated.

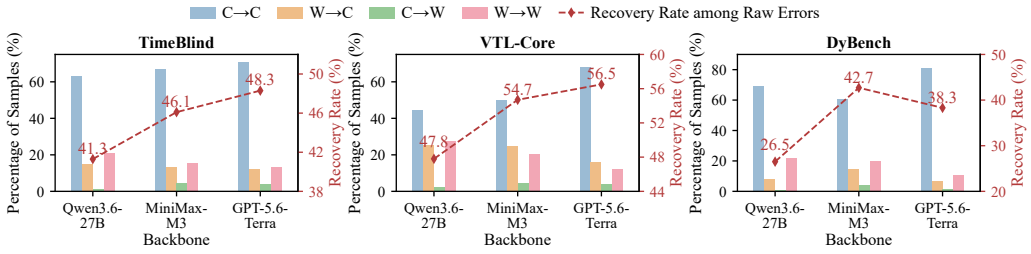

Figure 2: Correctness-transition analysis across backbones on TimeBlind, VTL-Core, and DyBench. Bars report the percentage of evaluation samples in each correctness transition from the sampled-frames baseline to SITE, where C and W denote correct and incorrect predictions, respectively. The dashed line reports the recovery rate among Sampled Frames errors, defined as the percentage of samples that are answered incorrectly under Sampled Frames baseline but are corrected by SITE, computed as: $\mathrm{W}{\to}\mathrm{C}/(\mathrm{W}{\to}\mathrm{C} + \mathrm{W}{\to}\mathrm{W})$.

Robustness Across Frame Sampling Strategies. To examine whether the effectiveness of SITE depends on a particular frame sampling strategy, we evaluate it with MiniMax-M3 under uniform sampling and Adaptive Keyframe Sampling (AKS) (Tang et al., 2025), using frame budgets of 8, 16, and 32. Given a frame budget $K \in \{8, 16, 32\}$, AKS selects up to K frames by jointly considering question relevance and video coverage, and presents them to the model in chronological order. Base interface directly answers from the selected frames, while SITE applies the full evidence-surfacing interface to the same sampled observations. As shown in Fig. 3, SITE consistently improves over the corresponding baseline across all evaluated frame budgets under both uniform sampling and AKS.

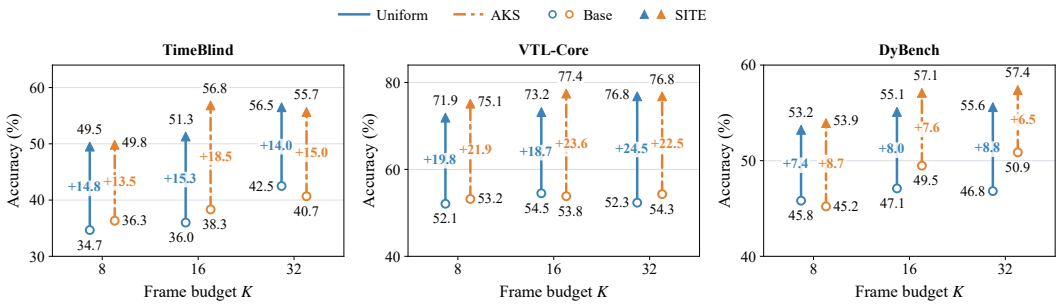

Figure 3: Comparison of Base and SITE accuracy across frame budgets $K \in \{8, 16, 32\}$ on TimeBlind, VTL-Core, and DyBench using uniform and AKS sampling. The reported metrics are I-Acc for TimeBlind, Acc for VTL-Core, and P-Acc for DyBench. Open circles and filled triangles denote Base and SITE accuracy, respectively. Each dashed connector links the two results obtained with the same sampling strategy and frame budget, with the annotated value indicating their absolute accuracy difference.

In addition, with either sampling strategy, SITE at eight frames outperforms the corresponding Base at 32 frames on all three datasets. This suggests that providing more frames alone does not ensure effective use of temporal information: how the available evidence is presented matters. The improvements under both uniform sampling and AKS further show that the benefits of SITE extend beyond a single frame-selection strategy.

Accuracy and Token Efficiency. We evaluate the inference efficiency of SITE by examining the relationship between token usage and temporal reasoning accuracy. Fig. 4 plots TimeBlind I-Acc against the tokens per question for MiniMax-M3. The stages of SITE provide several operating points with different accuracy–cost trade-offs. Within SITE, Route-Specific View improves I-Acc from 36.0% to 48.2%, while increasing token usage by only 20%. This improvement highlights the benefit of explicitly presenting the temporal evidence needed to answer the question.

We further compare SITE with TimeLogic, the strongest alternative in our main results. The Route-Specific View achieves an I-Acc of 48.2% using 3,654 tokens, closely matching TimeLogic's 48.7% with 8,403 tokens while reducing token usage by a factor of 2.3. At comparable accuracy, explicitly organizing relevant evidence therefore requires substantially fewer tokens than adaptively searching for additional informative frames. Adding Corroboration further raises I-Acc to 51.3% and outperforms TimeLogic by 2.6 percentage points. Overall, Route-Specific View offers comparable accuracy at substantially lower cost, while Corroboration trades additional computation for higher accuracy.

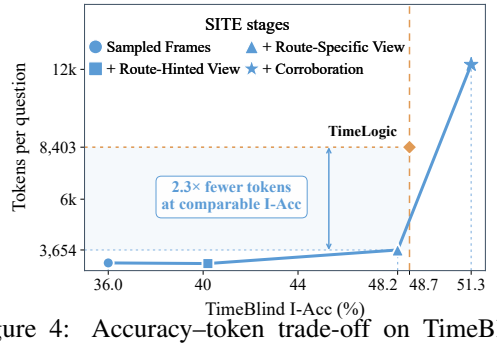

Figure 4: Accuracy–token trade-off on TimeBlind with MiniMax-M3. Blue markers connected by a solid line denote the successive stages of SITE. The orange diamond and dashed guides denote TimeLogic. Token cost includes all input and output tokens per question across all model calls, including the router and disagreement verifier.

5 Conclusion

We presented SITE, a training-free framework that routes each temporally-discriminative question to the required form of inter-frame evidence, constructs corresponding evidence views, and corroborates predictions across views while keeping the VLM frozen. Across three benchmarks and three VLM backbones, SITE consistently improves over Sampled frames and outperforms matched-backbone training-free baselines on every reported metric. It also remains competitive with post-trained temporal reasoning models. These results quantify an interface-recoverable component of temporal QA error and show that explicitly presenting question-relevant inter-frame relations can improve frozen-VLM inference.

AI USE STATEMENT

We used generative AI tools to polish the language of this paper and to assist with searching for and summarizing related literature. The writing assistance focused on improving phrasing, grammar, and clarity in text drafted by the authors. All AI-assisted content was reviewed and edited by the authors to ensure that it accurately reflects the work presented in this paper. The authors take full responsibility for the final manuscript, including its claims, results, and references.

REFERENCES

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, et al. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Tvbench: Redesigning video-language evaluation. 2024.

Dazhao Du, Jian Liu, Jialong Qin, Tao Han, Bohai Gu, Fangqi Zhu, Yujia Zhang, Eric Liu, Xi Chen, and Song Guo. Learning spatiotemporal sensitivity in video llms via counterfactual reinforcement learning. *arXiv preprint arXiv:2605.21988*, 2026.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026.

Zhenkun Gao, Xuhong Wang, Xin Tan, and Yuan Xie. Tpru: Advancing temporal and procedural understanding in large multimodal models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14953–14962. IEEE, 2023.

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

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14271–14280, 2024a. doi: 10.1109/CVPR52733.2024.01353.

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

Yeonkyung Lee, Dayun Ju, Youngmin Kim, Seil Kang, and Seong Jae Hwang. Vikey: Enhancing temporal understanding in videos via visual prompting. *arXiv preprint arXiv:2603.23186*, 2026.

Baiqi Li, Kangyi Zhao, Ce Zhang, Chancharik Mitra, Jean de Dieu Nyandwi, and Gedas Bertasius. Timeblind: A spatio-temporal compositionality benchmark for video llms. *arXiv preprint arXiv:2602.00288*, 2026.

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

Shicheng Li, Lei Li, Yi Liu, Shuhuai Ren, Yuanxin Liu, Rundong Gao, Xu Sun, and Lu Hou. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. In *European Conference on Computer Vision*, pp. 331–348. Springer, 2024b.

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

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

Wenqi Liu, Yunxiao Wang, Shijie Ma, Meng Liu, Qile Su, Tianke Zhang, Haonan Fan, Changyi Liu, Kaiyu Jiang, Jiankang Chen, et al. Videotemp-o3: Harmonizing temporal grounding and video understanding in agentic thinking-with-videos. *arXiv preprint arXiv:2602.07801*, 2026.

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

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

MiniMax. MiniMax-M3, 2026. URL `https://huggingface.co/MiniMaxAI/MiniMax-M3`. Model card.

OpenAI. GPT-5.6 system card, July 2026. URL `https://deploymentsafety.openai.com/gpt-5-6`.

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning. *arXiv preprint arXiv:2402.11435*, 2024.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026. URL `https://qwen.ai/blog?id=qwen3.6-27b`.

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

Sahil Shah, SP Sharan, Harsh Goel, Minkyu Choi, Mustafa Munir, Manvik Pasula, Radu Marculescu, and Sandeep Chinchali. Neus-qa: Grounding long-form video understanding in temporal logic and neuro-symbolic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8805–8813, 2026.

Chuyi Shang, Amos You, Sanjay Subramanian, Trevor Darrell, and Roei Herzig. Traveler: A modular multi-lmm agent framework for video question-answering. *arXiv preprint arXiv:2404.01476*, 2024.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. In *International Conference on Learning Representations*, volume 2025, pp. 7593–7734, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11854–11864. IEEE, 2023.

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

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

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

Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *European Conference on Computer Vision*, pp. 36–54. Springer, 2024c.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3272–3282. IEEE, 2025.

Zhaoyang Xu, Xusheng He, Wei Liu, Zhenyang Li, and Jianlong Wu. Timelogic challenge@ cvpr 2026: Strong mllms meet evidence-seeking agents for temporal-logic video question answering. *arXiv preprint arXiv:2606.01631*, 2026.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-voyager: Learning to query frames for video large language models. In *International Conference on Learning Representations*, volume 2025, pp. 84154–84179, 2025.

Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*, 2024.

A METHOD DETAILS

We describe the routing criteria, evidence view construction, and answer-selection rules of SITE in the order of its three stages. The view descriptions use the terminology of Sec. 3.

A.1 EVIDENCE ROUTING

Routing is performed by the frozen VLM backbone used for video question answering. The routing call contains only the question and a fixed prompt defining five routes. *Kinematic* questions concern object motion, including direction, speed, and motion state. *Event Order* questions concern the order, duration, or overlap of events. *Transition* questions concern changes in scene or object states. *Counterfactual* questions ask about outcomes under a stated intervention or changed condition. *Other* covers questions outside these four categories.

The prompt instructs the model to select exactly one route based on the temporal relation required by the question and return only its label. This classification is zero-shot, without labeled examples or parameter updates. No video frames, answer choices, or benchmark identity are provided. The predicted route determines the evidence views constructed by **Surface**. Questions assigned to *Other* use the route-hinted and timestamped views.

Routing prompt. The router selects the route that best captures the temporal relation required to answer the question. The complete prompt is shown below, where `<question>` is replaced with the input question.

```
You are routing module for a video temporal reasoning
system.  Read the question and determine what temporal
evidence is needed to answer it.  Choose exactly one route
according to the definitions below.

Kinematic:  The answer depends on how an entity moves over
time.  It requires examining motion properties such as
direction, speed, rotation, intensity, or repetition, or
distinguishing movement from a stationary state.  The
central question concerns the motion itself.

Event-Order:  The answer depends on when events occur or how
they relate in time.  It requires locating the relevant
events and comparing their order, duration, or overlap.
This includes determining which event happens first, what
follows another event, or whether events occur
simultaneously.

Transition:  The answer depends on how an object or scene
changes from one state to another.  It requires identifying
the states before and after a change and determining what
changed between them.  Examples include changes in
visibility, presence, or configuration, such as an object
appearing or a door opening.  The central question concerns
the resulting state change rather than the motion itself.

Counterfactual:  The answer depends on what would happen
under a hypothetical change to the situation shown in the
video.  It requires understanding the observed events and
reasoning about the consequences of an intervention, such as
removing an object, replacing an action, or changing a
condition.

Other:  The evidence needed to answer the question is not
adequately described by any of the four routes above.

If multiple routes appear relevant, select the one that
captures the relation directly required by the question.
Actions or objects mentioned only as context should not
determine the route.
```

```
Return exactly one of these route names: Kinematic,
Event-Order, Transition, Counterfactual, Other. Do not
include an explanation.

Question: <question>

Route:
```

A.2 Evidence Surfacing

Surface specifies the visual input and instruction for each view while retaining the original question and answer choices. The examples use 16 uniformly sampled frames resized to a maximum side length of 448 pixels. Previews show four images from each sequence. Each example pairs the visual input with the corresponding prompt excerpt.

Route-hinted view. This view retains the sampled frames and adds the selected question type to the prompt as a reasoning focus. Figure 5(b) shows the question-type label added to the prompt for a scene-change question.

Timestamped view. This view presents sampled frames with the frame index and video time in seconds overlaid on each image. These annotations are also present in *Sampled Frames* and the route-hinted view. The timestamped view additionally lists the frame times in the prompt to make their temporal positions explicit. Figure 5(c) illustrates the annotated frames and the corresponding prompt.

Transition view. This view preserves the sampled frames and changes only the prompt. The instruction asks the VLM to identify the initial and final states from the first and last quarters of the sequence and locate visual evidence of the change. Figure 6(a) illustrates this instruction for comparing two opposite scene transitions.

Chronological grid. This view arranges the same 16 frames into one 4×4 image with 300×300-pixel cells. The instruction specifies left-to-right and top-to-bottom order and applies the initial-to-final comparison to the grid. Figure 6(b) shows the composite input and its corresponding prompt.

Question-selected ordering view. The *Event Order* route pairs timestamped frames with an ordering instruction selected for the question. For the first-occurrence example in Fig. 7(a), the visual input remains unchanged. The prompt asks the VLM to locate the earliest visible occurrence of each candidate action and select the one that begins first.

Counterfactual view. This view keeps the observed frames unchanged and modifies the prompt to ground the relevant objects and events before reasoning about the stated intervention. Figure 7(b) illustrates an object-removal question.

Optical-flow view. The *Kinematic* route uses SEA-RAFT-M (Wang et al., 2024c) to analyze the full video and augments the sampled frames with motion-annotated frames and a summary of motion magnitude and direction over time. After reducing global translation, motion peaks guide the selection of additional frames whose moving regions receive boxes and direction arrows. The prompt asks the VLM to track the queried object, check whether the annotations refer to it, and infer its displacement. Figure 8 shows the sampled frames, additional visual inputs, and prompt for a motion-direction question. Implementation details are provided in Sec. A.4.

Question. What's the right option for how the scenes in the video change?
A. From the creek to the stairs. B. From the stairs to the creek.

Visual input shared by all three conditions

16 separate images in chronological order; 4 shown



Image 1 Image 6 Image 11 Image 16

(a) Sampled Frames

Answer from the sampled chronological video frames. Do not assume an event unless it is visually supported.

(b) Route-hinted view

Question type: *Transition*. Use the sampled chronological frames with this route as the reasoning focus, then answer directly.

(c) Timestamped view

Frames are chronological. Timestamps in seconds: 0.0, 1.3, 2.63, 3.97, 5.3, 6.63, 7.97, 9.3, 10.63, 11.97, 13.3, 14.63, 15.97, 17.3, 18.63, 19.97.

Figure 5: The route-hinted and timestamped views alongside *Sampled Frames*.

Question. What's the right option for how the scenes in the video change?
A. From the creek to the stairs. B. From the stairs to the creek.

(a) Transition view

16 separate images; 4 shown



Image 1 Image 6



Image 11 Image 16

The frames are chronological and labeled with timestamps. Do not solve this as generic recognition. First infer:
`INITIAL_SCENE`: dominant place/context in the first quarter of frames.
`FINAL_SCENE`: dominant place/context in the last quarter of frames.
`TRANSITION_EVIDENCE`: visual cues that show the change from initial to final.
Then compare every option as an ordered pair `<initial scene>` → `<final scene>`.

(b) Chronological grid

One composite image containing all 16 frames



The attached image is a 4x4 grid of video frames in chronological order from left to right and top to bottom. Each frame is labeled with index and timestamp.
`INITIAL_SCENE`: dominant place/context in the first quarter of the grid.
`FINAL_SCENE`: dominant place/context in the last quarter of the grid.
`TRANSITION_EVIDENCE`: visual cues that show the change from initial to final.

Figure 6: The transition view and chronological grid for a scene-change question.

(a) Question-selected ordering view

Question. What did the person do first?
A. Sat on the floor.
B. Put down the pillow.
Visual input. 16 sampled frames from 0–6 seconds.

16 separate images; 4 shown



Image 1



Image 6

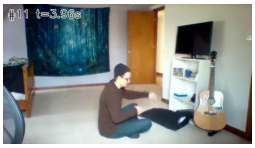

Image 11

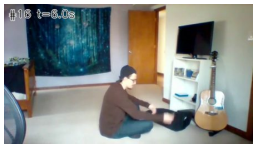

Image 16

For each answer option, find the earliest timestamp where that candidate action is visually supported. Choose the option whose action begins first in the shown video/window. Do not use common-sense scripts; use only the frame order. If an option phrase appears to be a result state, use the first frame where that state/action becomes visually true.

(b) Counterfactual view

Question. Which event will happen if the sphere is removed?
A. The yellow object and the brown cylinder collide.
B. The brown cylinder and the cube collide.
C. The yellow cylinder collides with the cube.
Visual input. 16 sampled frames.

16 separate images; 4 shown

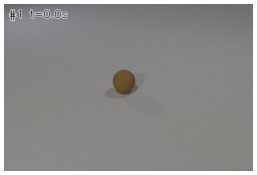

Image 1

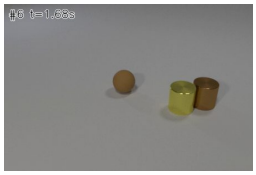

Image 6

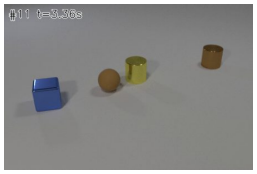

Image 11

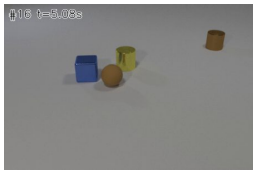

Image 16

First identify the relevant objects/events in the observed video. Then reason about the requested intervention or changed condition as a counterfactual event outcome. Choose the option whose outcome is most causally consistent with the observed dynamics and state transitions.

Figure 7: Examples of the question-selected ordering view and counterfactual view.

Question. Which direction does the blue sphere move in the video?
A. Down and to the right. B. Down and to the left.
C. Up and to the left. D. Up and to the right.

Sampled frames

16 separate images in chronological order; 4 shown

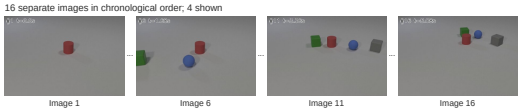

Image 1 Image 6 Image 11 Image 16

Additional motion-annotated frames

6 additional motion-annotated images; 2 shown

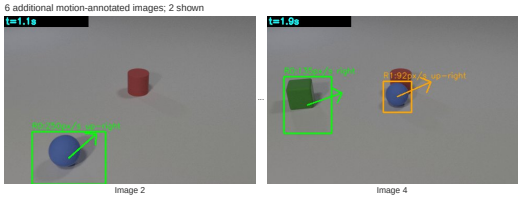

Image 2 Image 4

Motion summary excerpt

```
0.5-1.0s:  speed=3.7 [still]
1.0-1.5s:  speed=37.2 [fast]
dir=up-right (PEAK)
1.5-2.0s:  speed=20.9
dir=right
```

Prompt excerpt

Track ONLY this target object: blue sphere. Ignore other moving objects. Use color, shape, and relative position to keep the same object identity across frames.

Optical-flow evidence is attached as auxiliary evidence; use it only when the arrows/boxes correspond to the target object. If the flow evidence appears to describe a different object, trust the target's position change in the chronological frames.

```
TARGET_TRACK:
<first/middle/last visible
position of the target>
FLOW_CHECK:
<whether flow matches the target>
DISPLACEMENT:
<stationary | up-left | up-right |
down-left | down-right | left |
right | up | down>
```

Figure 8: An optical-flow view for a motion-direction question. Speed values denote normalized image motion rather than physical speed.

A.3 CROSS-VIEW CORROBORATION

Corroborate aggregates the frozen VLM's independent predictions across the evidence views constructed in Surface stage. If all predictions agree, it returns their shared answer. Otherwise, the VLM generates one additional prediction using the route-specific evidence view (Sec. 3.3), without access to earlier answers. Each prediction contributes one vote, and an answer with a unique highest vote count is returned. Only a tie for the highest count triggers the disagreement verifier, which receives the sampled frames, original evidence views, question, answer choices, and all preceding predictions. The verifier evaluates the competing answers against the visual evidence and returns the final answer directly, without adding its prediction to the vote count.

Disagreement verifier. The verifier uses the same frozen VLM backbone. It receives the chronological sampled frames, their timestamps, the original question and choices, the selected route, and the preceding candidate answers. Each candidate entry contains its evidence type identifying the corresponding evidence view, and answer label. The original visual input and instruction for every evidence view are also provided. The earlier reasoning traces are not included. The prompt instructs the VLM to treat candidates as hypotheses, check them against the visible temporal evidence, and return an answer from the original answer space as `FINAL: <answer>`.

Verifier prompt. The following template summarizes the verification instruction and input format. Each candidate's metadata occupies one line. Placeholders are replaced with the corresponding inputs.

```
A fixed video temporal reasoning pipeline produced several
candidate answers.  Use the candidates only as diagnostic
hypotheses; inspect the chronological frames and the
provided evidence views, including their visual inputs and
instructions, and choose the answer best supported by the
visible evidence.

route:  <route>
Candidate pool:
- evidence=<evidence view> | answer=<label>
<additional candidate entries>

Evidence views:
<view identifier>:  <original visual input and instruction>
<additional evidence views>

Resolve disagreements by checking the requested temporal
relation, motion, transition, or counterfactual outcome
against the video.  Do not average labels blindly; reject a
candidate if its evidence does not match the frames.

<question and answer choices>
<answer format instruction>
Frames are chronological.  Timestamps in seconds:
<timestamps>
<temporal window instruction, if applicable>

Output exactly:
FINAL: <answer>
```

A.4 OPTICAL FLOW IMPLEMENTATION

Evidence construction uses only the video and requires no VLM calls. SEA-RAFT-M estimates optical flow across the full video. Global translation is estimated by the componentwise median flow vector and subtracted. For residual flow magnitude ρ, normalized speed is $s = 100P_{95}(\rho)/(\Delta t D)$. Here P_q denotes the qth percentile, Δt is the actual frame interval and D is the resized image diagonal. Table 5 summarizes the inference and evidence construction settings.

Up to five detected motion peaks are retained by speed, or up to five uniform positions are used when no peaks are detected. The global maximum is included if absent. If the timeline contains at least four intervals but fewer than four positions are selected, four uniform positions are added. The

positions are deduplicated and sorted chronologically, yielding at most seven additional annotated frames.

Within each selected temporal window, the frame pair with the largest $P_{95}(\rho)$ supplies the residual flow and the frame to annotate. Thresholding and connected component analysis identify moving regions. Each region's velocity is the mean residual vector over pixels at or above its 70th magnitude percentile, divided by Δt. Retained regions receive bounding boxes and direction arrows. The textual summary reports the maximum normalized speed in each temporal bin and omits directions for bins marked `still`.

Table 5: Settings for optical-flow evidence construction. L denotes the number of motion intervals, A the region area and v the region speed in pixels per second.

Component	**Setting**
Flow inference	4 iterations; longest side ≤ 512 pixels; inference scale $1/2$; flow rescaled to resized frame coordinates.
Motion timeline	≈ 0.13 s stride; uniform subsampling to ≤ 80 pairs if needed.
Peak detection	Prominence $0.2(s_{\max} - s_{\min} + 10^{-9})$ and minimum separation $\max(1, \lfloor L/8 \rfloor)$ intervals; no extra suppression.
Region window	$[t - 0.2, t + 0.3]$ s clipped to video bounds; ≈ 0.15 s stride.
Motion mask	Threshold $\max(0.004D, 1.8\,\mathrm{median}(\rho))$; fallback $\max(0.5P_{90}(\rho), 1)$ if no foreground component. Opening 3×3; dilation 7×7.
Region selection	Area $\geq 0.3\%$ of the image; top 3 regions by $A\max(v, 1)$.
Direction labels	8 sectors; boundaries $22.5^\circ + 45^\circ k$ clockwise from screen right.
Region arrows	Mean residual-flow direction rounded to 1°; length $\min(70, 12 + v)$ pixels.
Motion summary	10 equal temporal bins; `fast` $\geq 60\%$ and `still` $< 25\%$ of the maximum reported speed.

B Additional Experiments

B.1 Comparison with Self-Consistency

We compare SITE with self-consistency (SC) under the same call budget to evaluate the benefit of using different evidence views rather than repeatedly answering from the same sampled frames. We conduct this comparison with MiniMax-M3 on TimeBlind. SC independently samples multiple answers at a temperature of 0.7 using the same Sampled Frames input and original answering prompt. The number of answers sampled for each question matches the number of calls that full SITE uses to generate answers. The most frequent answer is selected and ties are broken at random. This control increases inference effort without introducing evidence views tailored to the question.

Table 6 shows that SC raises Acc from 73.0% to 74.1% and I-Acc from 36.0% to 42.7% over Sampled Frames. Full SITE outperforms SC on all four metrics under the same call budget and achieves 78.6% Acc and 51.3% I-Acc. Its advantage over SC is largest in I-Acc. These results suggest that repeated sampling alone does not explain SITE's gains and support the value of evidence presentation tailored to the question.

Table 6: Comparison with self-consistency on TimeBlind with MiniMax-M3. SC and full SITE use the same call budget for each question.

Method	**Acc (%)**	**Q-Acc (%)**	**V-Acc (%)**	**I-Acc (%)**
Sampled Frames (Base)	73.0	53.3	57.6	36.0
Self-consistency (SC)	74.1	55.7	62.1	42.7
Full SITE	**78.6**	**62.1**	**68.7**	**51.3**

B.2 Comparison with Random Routing

To assess the contribution of question-conditioned routing in SITE, we replace SITE's semantic router with randomized routing policies while keeping the sampled frames, evidence construction, corroboration procedure, and inference configuration unchanged. Table 7 indicates that indiscriminately constructing alternative evidence views provides limited benefit; effective evidence surfacing

requires the constructed view to match the temporal relation queried by the question. This further highlights the importance of explicitly surfacing question-relevant evidence.

Table 7: Effect of question-conditioned routing in SITE on TimeBlind with MiniMax-M3.

Method	**Acc (%)**	**Q-Acc (%)**	**V-Acc (%)**	**I-Acc (%)**
Sampled Frames (Base)	73.0	53.3	57.6	36.0
Random SITE	75.2	56.4	61.3	45.0
Full SITE	**78.6**	**62.1**	**68.7**	**51.3**

C LIMITATIONS

SITE is designed to recover temporal reasoning that remains dormant under the standard presentation, rather than to expand the capability of the frozen VLM. When the backbone lacks the perceptual or reasoning capabilities required for the task, making temporal evidence explicit may still be insufficient to produce a correct answer. Full SITE also requires several VLM calls per question and therefore costs several times more than single-call inference on average, although this overhead is comparable to other multi-call inference methods.