VIDEOMEND: MENDING TEMPORAL CONTINUITY IN VIDEO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video reasoning requires models to capture fine-grained temporal dynamics, including action directionality, motion dynamics, and state transitions. Yet, constrained by computational costs, many Video Large Language Models resort to sparse frame sampling, which fragments the temporal stream into isolated visual snapshots and encourages single-frame bias: models derive answers from salient objects in individual frames rather than from the temporal evolution of events. We propose **VIDEOMEND**, a framework that explicitly reasons over inter-frame changes to mend the temporal discontinuity introduced by frame sampling. Its core is Ref-Delta Chain-of-Thought (Ref-Delta CoT), a structured reasoning format organized around two complementary signals: *Reference* anchors the static scene semantics at temporal anchor points, while *Delta* explicitly captures the changes between adjacent frames, encouraging the model to articulate temporal transitions rather than derive answers from individual frames in isolation. To train it, we adopt a two-stage pipeline. The SFT stage provides cold-start initialization with data specifically targeting temporal direction and speed perception, including antonymous semantic pairs that supervise temporal direction and symmetric constructions that decouple semantic complexity from speed. The RL stage further enforces the visual faithfulness of Deltas via Delta-Mask GRPO (DM-GRPO), which constructs counterfactual inputs via delta frame masking and uses a judge model to provide an explicit training signal for the visual grounding of Deltas, ensuring inter-frame changes are grounded in visual observation rather than language priors. Experiments on six benchmarks show consistent gains over strong baselines, most notably on TempCompass, reflecting **VIDEOMEND**'s capacity to capture temporal continuity.

1 INTRODUCTION

Effective video reasoning requires Multimodal Large Language Models (MLLMs) to capture temporal dynamics, including events, interactions, and outcomes, beyond static frame recognition. Despite advances in Video Large Language Models Cheng et al. (2024); Zhang et al. (2024); Bai et al. (2025) and reinforcement learning–based approaches Feng et al. (2025); Yan et al. (2025); Maaz et al. (2025), modeling temporal continuity remains challenging.

Computational constraints lead many Video LLMs to sample sparse frames, leaving gaps in temporal evidence Hu et al. (2025a); Zhang et al. (2025c); Hu et al. (2025c); Zhang et al. (2025b). This can encourage **single-frame bias**: predictions rely on salient objects or transient states in individual frames rather than event evolution, as illustrated on the left of Figure 1. Yet action direction, motion speed, and state transitions depend on relationships across frames. The central challenge is therefore to reason about continuous events from discrete observations.

Faced with this dilemma, prior work has recognized temporal modeling as a key challenge in video reasoning Liu et al. (2024b); Huang et al. (2024); Ren et al. (2024); Yang et al. (2025). Recent reinforcement learning approaches attempt to encourage the use of temporal continuity through reward design Feng et al. (2025); Wang et al. (2025a); Li et al. (2025b); Maaz et al. (2025); Dang et al. (2025); Luo et al. (2025). Yet their reasoning processes still treat sampled frames as independent observations, without sufficiently leveraging what changes between frames in the reasoning chain. The challenge, therefore, is not simply whether models attend to temporal continuity, but how to

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.

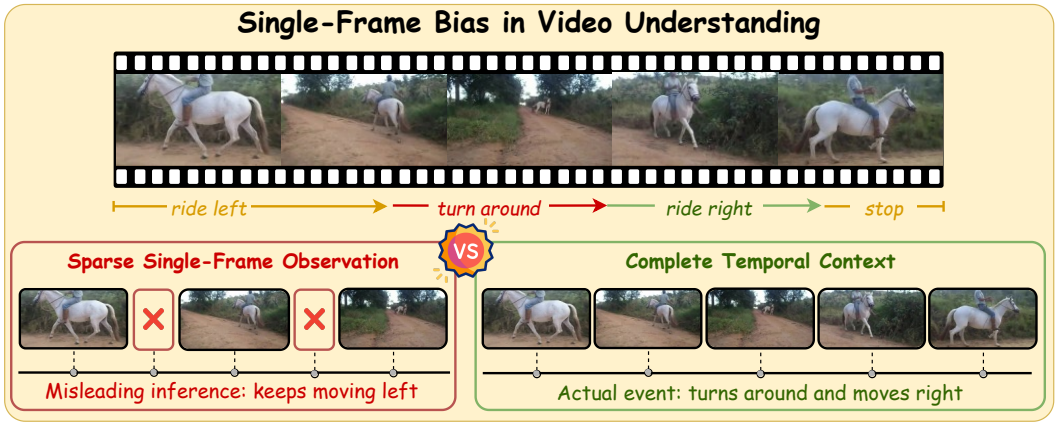


Figure 1: Illustration of single-frame bias. Sparse observations suggest continued leftward motion (left), while fuller temporal context reveals a reversal to rightward motion (right).

reason over the temporal gaps between them. We argue that the reasoning structure itself should model what happens between frames, capturing the missing temporal continuity underlying discrete samples. This motivates a shift from *sample-then-understand* to *sample-then-reason-then-understand*.

Building on this insight, we propose **VIDEOMEND**, a framework that explicitly reasons over inter-frame changes to mend the temporal discontinuity introduced by frame sampling. We introduce Ref-Delta Chain-of-Thought (Ref-Delta CoT), which organizes video reasoning around two complementary signals: *Reference* anchors the static semantics of each sampled frame, while *Delta* explicitly captures the changes between adjacent frames, as illustrated on the right side of Figure 1. This design encourages the model to articulate transitions between frames, rather than deriving answers from salient objects in individual frames.

We train **VIDEOMEND** in two stages using the Ref-Delta reasoning structure. We first construct targeted training data for temporal direction and motion speed, two signals that are difficult to infer from individual frames but essential for generating faithful Deltas. Specifically, we construct antonymous semantic pairs within similar scene contexts to supervise temporal direction, and use symmetric constructions that decouple semantic complexity from motion speed to supervise speed perception. These data are annotated in the Ref-Delta format and used in the SFT stage for cold-start initialization, teaching the model to organize reasoning around References and Deltas. We then further optimize the model in the RL stage with Delta-Mask GRPO (DM-GRPO). Since SFT alone may still allow the model to produce plausible Deltas from language priors or rely on single-frame shortcuts, DM-GRPO leverages judge-annotated temporal segments and frame masking to construct counterfactual inputs. By evaluating whether predictions change appropriately under such counterfactual masking, DM-GRPO provides an explicit training signal for the visual grounding of Deltas. Together, targeted temporal supervision, SFT cold-start initialization, and DM-GRPO enable **VIDEOMEND** to better capture the temporal continuity of video by explicitly reasoning over inter-frame changes. Our contributions are summarized as follows:

- We propose Ref-Delta Chain-of-Thought. Ref-Delta CoT uses References to anchor static scene semantics of sampled frames and Deltas to explicitly capture adjacent-frame changes, mending the temporal discontinuity caused by sparse frame sampling and mitigating single-frame bias.
- We build a two-stage training pipeline with Delta-Mask GRPO. The SFT stage uses antonymous semantic pairs and symmetric speed constructions to supervise temporal direction and motion speed. The RL stage introduces DM-GRPO, which constructs counterfactual inputs via delta-frame masking and uses Judge signals to enforce Delta visual faithfulness.
- **VIDEOMEND** improves over the base model on six benchmarks and raises TempCompass from 72.6% to 76.7%, validating inter-frame change modeling for temporal reasoning.

2 Related Works

Video Large Language Models. Recent Video LLMs have substantially advanced video understanding through video-text alignment, instruction tuning, and improved spatiotemporal modeling Maaz et al. (2024); Cheng et al. (2024); Zhang et al. (2024); Bai et al. (2025); Chen et al. (2024b); Zhang et al. (2025a). Due to visual-token costs, however, most models operate on sparsely sampled frames, motivating frame-selection methods that seek more informative observations under limited budgets Hu et al. (2025a); Zhang et al. (2025c); Hu et al. (2025c); Zhang et al. (2025b). While these methods improve *which frames to observe*, inter-frame changes remain largely implicit in the subsequent reasoning process. VideoMend instead targets this reasoning-level discontinuity by explicitly representing changes between sampled frames.

Reinforcement Learning for Video Reasoning. Recent works adapt RL-based reasoning to video understanding through temporal, grounding, and spatiotemporal rewards Feng et al. (2025); Wang et al. (2025a); Li et al. (2025b); Yan et al. (2025); Li et al. (2025c); Park et al. (2025); Yang et al. (2026); Yu et al. (2025). These methods demonstrate that RL can improve video reasoning and visual grounding, but typically optimize reasoning over sampled visual observations without explicitly structuring the reasoning trace around inter-frame changes. In contrast, VideoMend introduces Ref-Delta CoT to represent such changes explicitly and DM-GRPO to enforce their grounding through counterfactual frame masking. We provide a more detailed discussion of related methods in Appendix A.

3 Method

Frame sampling reduces a continuous video $\mathcal{V} = \{v(t)\}_{t=0}^{T}$ to discrete observations $F = \{f_i = v(t_i)\}_{i=1}^{N}$, leaving inter-frame dynamics without explicit representation. To address this discontinuity, we propose **VideoMend**, a framework that explicitly incorporates inter-frame changes into its reasoning structure. Its core is *Ref-Delta Chain-of-Thought (Ref-Delta CoT)*: *References* describe static visual states at selected observation times, while *Deltas* describe changes over temporal intervals that may span multiple sampled frames. Together, they connect discrete observations through explicit temporal transitions, encouraging reasoning about event evolution rather than isolated frame content. **VideoMend** follows a two-stage training pipeline. First, we construct **VideoMend-RefDelta-CoT-13K** from base video data and augmentation targeting temporal direction and motion speed, providing SFT cold-start initialization in the Ref-Delta format. We then construct **DM-Data-8K** and apply DM-GRPO, using counterfactual masking to encourage visually grounded Delta reasoning. The pipeline aims to mitigate single-frame bias by combining structured temporal supervision with counterfactual optimization.

3.1 Data Collection

Our training corpus comprises 13,248 video question–answer pairs from two complementary sources. The first source includes 10,392 samples drawn from nine public video benchmarks, covering diverse reasoning capabilities, including action recognition, causal reasoning, procedural understanding, spatiotemporal composition, and fine-grained perception. These data provide broad supervision for video reasoning but offer limited targeted supervision for temporal direction and motion speed, as these properties can often be partially inferred from static-frame semantics without explicitly modeling temporal changes. To address this limitation, we further construct 2,856 purpose-built augmentation samples, representing 21.6% of the corpus, to mitigate reliance on static semantic shortcuts and encourage reasoning about inter-frame changes. A source-by-source breakdown of the base and augmentation data is provided in Appendix B.1.

3.2 Data Augmentation

Faithful Deltas require complementary supervision on the *direction* and *speed* of change: direction distinguishes opposing state transitions, while speed characterizes how quickly they unfold. We construct two augmentation subsets that emphasize temporal evidence over static semantic shortcuts. Figure 2 illustrates the source pairs and their composition.

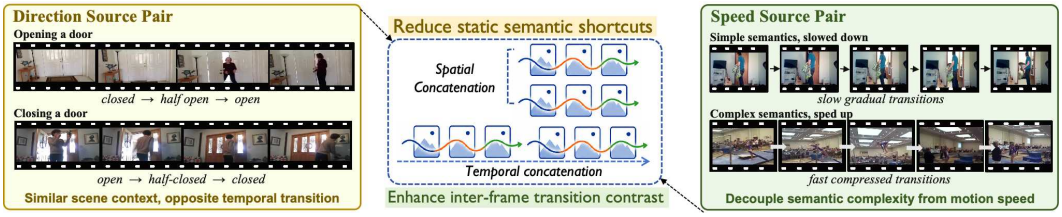


Figure 2: Temporal augmentation. Left: opposing actions. Center: spatial and temporal concatenation. Right: simple clips are slowed down and complex clips sped up to challenge complexity-based speed shortcuts.

Temporal Direction: Antonymous Action Pairs. We construct direction-sensitive QA pairs from Something-Something-V2 Goyal et al. (2017) using opposing actions. Figure 2 (left) illustrates opening and closing a door: similar scenes accompany opposite transitions, from closed to open and vice versa. For each base pair, we sample both actions and combine the clips spatially within a shared frame or temporally in sequence (center panel). Their similar scene context makes static appearance less discriminative, encouraging the model to track the direction of change. This yields 1,340 samples from 95 base pairs.

Temporal Speed: Complexity–Speed Counter-Pairing. Models may associate simple content with fast motion and complex content with slow motion. To challenge this potential shortcut, we slow down clips with few objects and simple actions and speed up clips with richer scenes or multi-step actions (Figure 2, right). The clips are combined spatially or temporally into speed-sensitive QA pairs, countering this complexity–speed association and encouraging comparison of temporal change rates. This yields 1,516 samples. Dataset composition is detailed in Appendix B.1.

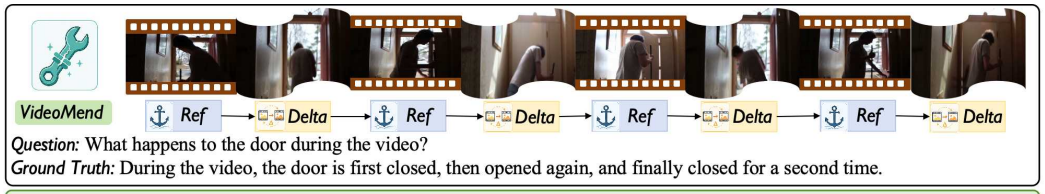


Figure 3: Illustration of the Ref-Delta CoT structure. References anchor sampled frames, while Deltas describe inter-frame changes to reconstruct temporal continuity.

3.3 Ref-Delta CoT Generation

To provide Ref-Delta CoT supervision, we construct **VideoMend-RefDelta-CoT-13K**. As shown in Figure 3, References capture static visual states, Deltas describe temporal changes, and a reasoning summary integrates Delta evidence to derive the answer. We use a vision-language annotator to jointly interpret visual states and describe question-relevant changes, grouping intervals by semantics rather than predefined segmentation rules.

Generation. We deploy Qwen3.5-122B-A10B locally via vLLM. Given the video, question, and ground-truth answer, it jointly generates Reference timestamps, Delta intervals, and descriptions

in one generation call, without a separate segmentation algorithm. The prompt includes an FPS constraint based on the actual sampling rate and one in-context example (Appendix B.2). Dimension-specific hints guide direction- or speed-relevant Deltas for the augmentation subsets in Section 3.2. Structural and semantic filtering, followed by judge-based answer-leakage checks, controls annotation quality (Appendix B.3). A human audit finds 98.6% of 1,000 retained traces correct, supporting supervision quality (Appendix C.3).

3.4 DELTA-MASK GRPO

SFT teaches the model to follow the Ref-Delta format but may not eliminate single-frame bias: plausible Deltas can coexist with answers driven by salient objects in individual frames, static descriptions, or language priors rather than inter-frame changes. To mitigate this shortcut, we propose **Delta-Mask GRPO (DM-GRPO)**, which masks frames in a selected Delta interval and uses an offline Judge to assess whether the remaining evidence is sufficient to answer the question (Figure 4). The resulting counterfactual signal is combined with per-response rewards to update the policy.

Delta-frame masking. For each training sample (F, q) with sampled frames $F = \{f_k\}_{k=1}^{N}$, we select one Delta interval $[s_d, e_d]$ from the offline Ref-Delta annotation z^*. Let t_k be the timestamp of frame f_k. The frames covered by this Delta are indexed by

$$\mathcal{I}_d = \{k : \lfloor t_k \rfloor \in [\lfloor s_d \rfloor, \lfloor e_d \rfloor]\}. \tag{1}$$

We construct the masked video $\tilde{F}$ by zeroing out the pixels of $\{f_k\}_{k \in \mathcal{I}_d}$ while preserving their temporal positions, so the input length and temporal structure remain unchanged. The policy π_θ is then sampled on both (F, q) and $(\tilde{F}, q)$, yielding the original group $\{o_i\}_{i=1}^{G}$ and the delta-masked group $\{\tilde{o}_j\}_{j=1}^{\tilde{G}}$. The latter is used only to estimate the counterfactual signal and does not participate in advantage computation or gradient updates. DM-Data-8K provides 14,895 mask patterns for this procedure, with their distribution shown in Figure 10.

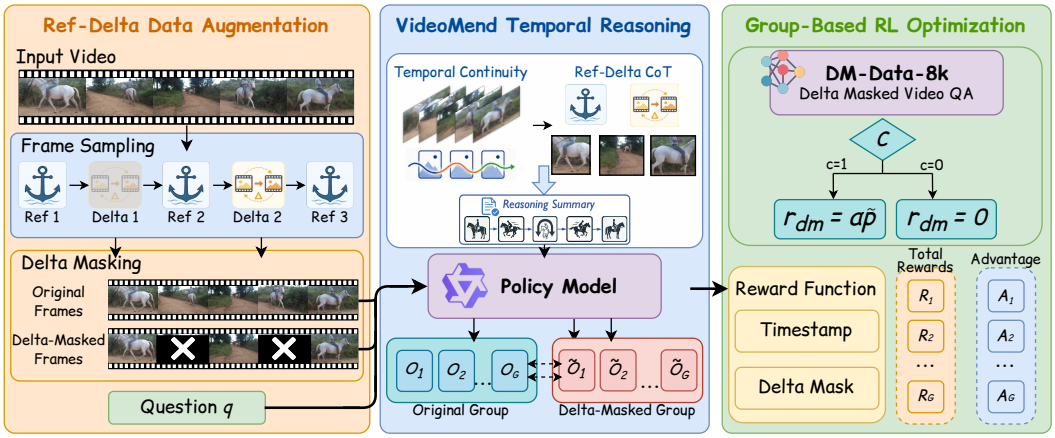


Figure 4: Overview of Delta-Mask GRPO. A selected Delta interval is masked and judged as critical or non-critical; the delta-masked group is used only to estimate the counterfactual Delta-Mask signal, while the original group receives the final rewards for policy updates.

Evidence-masking Judge. An interval is non-critical if masking its frames leaves sufficient visual evidence to answer the question. It may contain salient or redundant information, but the question remains answerable without it. We construct **DM-Data-8K** with 8,008 video–question pairs and 14,895 mask patterns (Appendix B.4). For each interval, we query Qwen3.5-122B-A10B on the masked input $(\tilde{F}, q)$ and use answer correctness to assess whether sufficient evidence remains. A correct answer yields $c = 1$ (non-critical); otherwise, $c = 0$, conservatively disabling the Delta-Mask term. A class-balanced human audit of 800 patterns shows 96.75% agreement with the Judge (Appendix C.2). Only $c = 1$ samples are eligible for a non-zero Delta-Mask term; $c = 0$ samples retain the base and timestamp rewards.

Reward design. DM-GRPO uses three signals: the core Delta-Mask counterfactual reward, which encourages robustness to the removal of non-critical Delta intervals; the per-response base reward,

which enforces answer correctness and format compliance; and the timestamp plausibility reward, which encourages temporally well-formed Ref-Delta CoT traces.

Delta-Mask reward. To quantify counterfactual robustness, let $\tilde{p}$ denote the proportion of correct responses in the delta-masked group:

$$\tilde{p} = \frac{1}{\tilde{G}} \sum_{j=1}^{\tilde{G}} \mathbb{I}[\tilde{o}_j \text{ answers } q \text{ correctly}]. \tag{2}$$

The Delta-Mask reward is defined as

$$r_{\mathrm{dm}} = \begin{cases} \alpha\tilde{p}, & \text{if } c = 1, \text{ non-critical,} \\ 0, & \text{if } c = 0, \text{ critical.} \end{cases} \tag{3}$$

where α is the reward scale and $c \in \{0, 1\}$ is the Judge-provided non-criticality label: $c = 1$ indicates that the masked Delta interval is non-critical, while $c = 0$ indicates that it is critical. The effect of r_{dm} depends on reliance on the masked non-critical interval: if the model uses salient but answer-irrelevant content as a shortcut, masking lowers $\tilde{p}$ and the reward; otherwise, correctness is preserved, yielding a larger $\tilde{p}$ and a stronger reward. Thus, r_{dm} discourages reliance on non-critical Delta evidence while preserving informative cues.

When $c = 0$, the masked interval is critical, so a performance drop does not indicate flawed reasoning. We therefore set $r_{\mathrm{dm}} = 0$ to keep the counterfactual test reward-neutral.

Base reward. Each original response o_i first receives a base reward

$$r_i^{\mathrm{base}} = r_{\mathrm{acc},i} + \lambda_{\mathrm{fmt}} r_{\mathrm{fmt},i}, \tag{4}$$

where $r_{\mathrm{acc},i} \in \{0, 1\}$ checks answer correctness and $r_{\mathrm{fmt},i} \in \{0, 1\}$ checks whether valid `<think>` and `<answer>` fields are present. If the format is invalid, we set $r_{\mathrm{fmt},i} = 0$.

Timestamp reward. In addition to the base reward, each original response also receives a timestamp plausibility reward $r_{\mathrm{ts},i} \in [0, 1]$ to encourage temporally well-formed Ref-Delta CoT traces. The timestamp reward first applies hard validity checks: timestamps must be parsable, each `<time_range>` (s, e) must satisfy $e \geq s$, and the maximum timestamp must not exceed the video duration by more than 1 second. If the response format is invalid or any timestamp validity check fails, $r_{\mathrm{ts},i} = 0$; otherwise,

$$r_{\mathrm{ts},i} = s_{\mathrm{mono},i} + s_{\mathrm{overlap},i} + s_{\mathrm{cov},i}. \tag{5}$$

Here, $s_{\mathrm{mono},i}$ checks whether interval start times are monotonic within a 1-second tolerance, $s_{\mathrm{overlap},i}$ checks whether adjacent intervals avoid overlaps longer than 3 seconds, and $s_{\mathrm{cov},i} = \mathrm{clip}((\mathcal{C}_i - C_{\min})/(C_{\max} - C_{\min}), 0, 1)$, where $\mathcal{C}_i = |\bigcup \text{time_ranges}_i|/T$ is the merged Delta coverage. The per-response reward before applying the Delta-Mask term is then

$$r_i = r_i^{\mathrm{base}} + \lambda_{\mathrm{ts}} r_{\mathrm{ts},i}. \tag{6}$$

Final reward and policy update. The Delta-Mask reward is applied only to original-group responses that answer correctly:

$$R_i = \begin{cases} r_i + r_{\mathrm{dm}}, & \text{if } o_i \text{ answers correctly,} \\ r_i, & \text{otherwise.} \end{cases} \tag{7}$$

The Delta-Mask reward is applied only to correct responses in the original group $\{o_i\}_{i=1}^{G}$, so that counterfactual robustness does not provide an additional bonus to incorrect original responses. Advantages are computed only within this group; the delta-masked group $\{\tilde{o}_j\}_{j=1}^{\tilde{G}}$ is used only to estimate the counterfactual signal and is excluded from policy-gradient updates.

Let A_i denote the advantage of response o_i, computed within the original response group. Let ϵ be the clipping coefficient, β the KL regularization coefficient, $\pi_{\theta_{\mathrm{old}}}$ the sampling policy, and π_{ref} the reference policy. We optimize

$$\mathcal{L}_{\text{DM-GRPO}} = -\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} \min\left(\rho_i A_i,\ \mathrm{clip}(\rho_i, 1-\epsilon, 1+\epsilon) A_i\right) - \beta D_{\mathrm{KL}}(\pi_\theta(\cdot|F,q) \,\|\, \pi_{\mathrm{ref}}(\cdot|F,q))\right], \tag{8}$$

where $\rho_i = \pi_\theta(o_i|F,q)/\pi_{\theta_{\mathrm{old}}}(o_i|F,q)$.

4 EXPERIMENTS

4.1 SETUP

Benchmarks and Metrics. Following prior video reasoning studies Feng et al. (2025); Wang et al. (2025a); Li et al. (2025b); Yan et al. (2025); Maaz et al. (2025), we evaluate **VIDEOMEND** on six public benchmarks: VSI-Bench, VideoMMMU, and MMVU for reasoning-intensive video understanding, and MVBench, TempCompass, and VideoMME for general video understanding and temporal perception Yang et al. (2025); Hu et al. (2025b); Zhao et al. (2025); Li et al. (2024a); Liu et al. (2024b); Fu et al. (2025). TempCompass is particularly aligned with our focus on temporal direction, event order, and motion speed. We report accuracy (%), using MMVU's multiple-choice subset and VideoMME without subtitles. All evaluations use 32 frames, greedy decoding, and up to 4,096 output tokens. Multiple-choice predictions are extracted from the `<answer>` field and scored using each benchmark's official protocol.

Implementation Details. We train **VIDEOMEND** from Qwen2.5-VL-7B Bai et al. (2025) in two stages, both with a learning rate of 1×10^{-6}. SFT runs for 1 epoch with ZeRO-2 on 8 A100 GPUs, using an effective batch size of 32 (1 sample per GPU, 4 gradient accumulation steps), bf16 precision, and Flash Attention 2. We uniformly sample up to 32 frames per video during SFT. RL uses DM-GRPO with ZeRO-3, allocating 7 GPUs to training and 1 to vLLM-based generation. For each training sample, we generate $G = 8$ responses from the original video and $\tilde{G} = 4$ from the delta-masked video at temperature 1.0. RL runs for 761 optimization steps with KL coefficient $\beta = 0.04$. We set the Delta-Mask reward scale to $\alpha = 0.3$, yielding $r_{\mathrm{dm}} = \alpha\tilde{p}$ for non-critical intervals and zero otherwise. This bonus is added only to correct original-video responses. The timestamp reward weight is $\lambda_{\mathrm{ts}} = 0.2$.

Table 1: Comparison with video MLLM baselines on video reasoning and general video benchmarks. Bold and underline denote the best and second-best RL-based results, respectively. Δ denotes the absolute improvement over Qwen2.5-VL-7B in percentage points.

Model	Video Reasoning Benchmarks			General Video Benchmarks		
	VSI-Bench	VideoMMMU	MMVU (mc)	MVBench	TempCompass	VideoMME w/o sub
Non-RL Video MLLMs						
GPT-4o Hurst et al. (2024)	34.0	61.2	75.4	–	–	71.9
LLaMA-VID Li et al. (2024b)$_{\text{ECCV'24}}$	–	–	–	41.9	45.6	–
VideoLLaMA2 Cheng et al. (2024)$_{\text{arXiv'24.06}}$	–	–	44.8	54.6	–	47.9
LongVA-7B Zhang et al. (2025b)$_{\text{TMLR'24}}$	29.2	23.9	–	–	56.9	52.6
VILA-1.5-8B Lin et al. (2024)$_{\text{CVPR'24}}$	28.9	20.8	–	–	58.8	–
VILA-1.5-40B Lin et al. (2024)$_{\text{CVPR'24}}$	31.2	34.0	–	–	–	60.1
Video-UTR-7B Yu et al. (2025)$_{\text{ICLR'25}}$	–	–	–	58.8	59.7	52.6
LLaVA-OneVision-7B Li et al. (2025a)$_{\text{TMLR'25}}$	32.4	33.8	49.2	56.7	–	58.2
Kangaroo-8B Liu et al. (2024a)$_{\text{arXiv'24.08}}$	–	–	–	61.1	62.5	56.0
Qwen2.5-VL-7B Bai et al. (2025)$_{\text{arXiv'25.02}}$	30.1	48.1	60.0	59.0	72.6	56.6
RL-based Video MLLMs						
Video-R1 Feng et al. (2025)$_{\text{NeurIPS'25}}$	35.8	49.8	63.8	63.9	73.2	59.3
VideoChat-R1 Li et al. (2025b)$_{\text{arXiv'25.04}}$	33.0	**52.0**	64.8	63.6	74.5	64.1
VideoChat-R1.5 Yan et al. (2025)$_{\text{NeurIPS'25}}$	36.1	50.0	67.0	65.7	73.9	<u>64.8</u>
VideoRFT Wang et al. (2025a)$_{\text{NeurIPS'25}}$	36.8	48.1	66.7	62.1	73.7	59.8
MOSS-ChatV Tao et al. (2025)$_{\text{arXiv'25.09}}$	–	50.2	–	<u>67.6</u>	72.9	60.0
Video-R2 Maaz et al. (2025)$_{\text{CVPR'26}}$	**39.4**	50.8	<u>67.4</u>	67.5	<u>74.9</u>	63.8
VIDEOMEND-7B	<u>37.4</u>	<u>51.2</u>	**67.6**	**67.9**	**76.7**	**65.1**
Δ vs. Qwen2.5-VL-7B	+7.3	+3.1	+7.6	+8.9	+4.1	+8.5

4.2 MAIN RESULTS

Table 1 compares **VIDEOMEND** with existing video MLLMs on six benchmarks spanning reasoning-intensive and general video understanding tasks. To broaden the evaluation, we additionally report results on AoTBench, TVBench, NExT-QA, and MECD in Appendices D.0.1, D.0.2, D.1.1, and D.1.2, respectively. These benchmarks probe arrow-of-time reasoning, fine-grained temporal understanding, causal and temporal video QA, and multi-event causal discovery. Inference throughput and latency are reported in Appendix D.2.

Table 2: Ablation study for VIDEOMEND.

Model	Video Reasoning			Video Understanding		
	VSI-Bench	VideoMMMU	MMVU	MVBench	TempCompass	VideoMME
Qwen2.5-VL-7B Bai et al. (2025)	30.1	48.1	60.0	59.0	72.6	56.6
w/o Ref–Delta Markers	36.0	50.5	66.5	66.2	74.8	63.8
w/o Data Augmentation	36.8	50.7	67.0	67.2	74.2	64.2
w/o DM-GRPO	33.2	49.5	64.5	64.2	73.3	61.1
w/o r_{ts}	37.1	50.8	67.3	67.7	76.1	64.8
w/o r_{dm}	35.8	50.2	66.2	65.9	73.5	62.3
VIDEOMEND	**37.4**	**51.2**	**67.6**	**67.9**	**76.7**	**65.1**

Consistent effectiveness. **VIDEOMEND** outperforms the Qwen2.5-VL-7B base model on every evaluated benchmark, raising the average score from 54.4% to 61.0%, a gain of 6.6 percentage points. Improvements extend across reasoning-intensive tasks and broader video understanding benchmarks Rasheed et al. (2025); Wu et al. (2024); Wang et al. (2025b). This consistency suggests that explicitly representing inter-frame changes with Ref-Delta CoT and strengthening their visual grounding with DM-GRPO are useful across different evaluation settings, rather than being tailored to a single benchmark.

Temporal reasoning gains. **VIDEOMEND** shows clear gains on TempCompass, improving Qwen2.5-VL-7B from 72.6% to 76.7% and outperforming the strongest prior RL-based baseline by 1.8%. Since TempCompass evaluates temporal direction, event order, and motion speed, this improvement directly reflects the benefit of modeling inter-frame dynamics. The result supports our motivation that explicitly modeling Deltas helps mitigate the temporal discontinuity introduced by sparse frame sampling.

Generalization ability. Beyond temporal-oriented evaluation, **VIDEOMEND** also performs strongly on broader video understanding benchmarks, including MMVU, MVBench, and VideoMME Zhao et al. (2025); Li et al. (2024a); Fu et al. (2025). While it is not the best on VSI-Bench and VideoMMMU, these benchmarks emphasize complex scene understanding, spatial reasoning, and domain-specific knowledge, which are less directly aligned with fine-grained temporal continuity. **VIDEOMEND** still consistently improves over the base model, showing that it enhances temporal modeling while preserving general video reasoning ability.

4.3 ABLATION STUDY

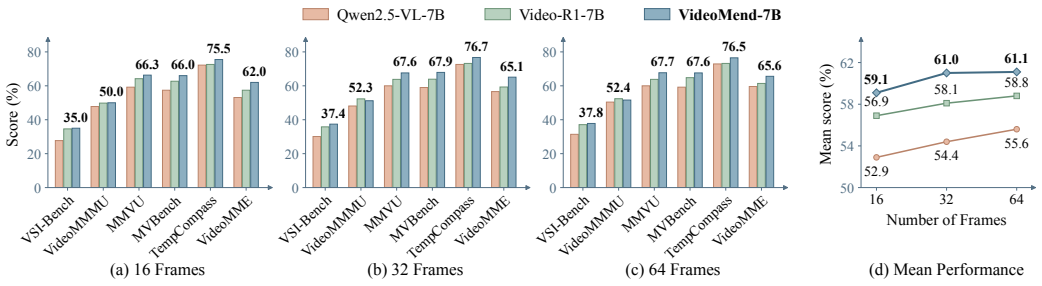


Figure 5: Performance at matched input budgets of 16, 32, and 64 frames. Panels (a)–(c) report per-benchmark accuracy, while panel (d) summarizes mean performance. VIDEOMEND consistently outperforms Qwen2.5-VL-7B and Video-R1 across all budgets, with diminishing returns beyond 32 frames.

Effects of Core Components. We ablate explicit Ref–Delta markers, temporal augmentation, and DM-GRPO (Table 2). Removing Reference/Delta markers from SFT traces while preserving reasoning content, answers, and temporal information reduces TempCompass accuracy from 76.7% to 74.8%, supporting the value of explicitly distinguishing states from changes. Removing direction

and speed augmentation lowers TempCompass accuracy to 74.2%, the largest drop for this variant across the six benchmarks. Omitting DM-GRPO yields SFT-only and causes the largest degradation among these component ablations: TempCompass, MVBench, and VideoMME fall to 73.3%, 64.2%, and 61.1%, respectively. Together, these results support the contributions of all three components.

Effects of Reward Design. To isolate counterfactual supervision and temporal regularization, we remove r_{dm} and r_{ts} separately while keeping other rewards unchanged (Table 2). Removing r_{dm} reduces accuracy across all six benchmarks by 1.0–3.2 percentage points, with the largest drops on TempCompass (3.2) and VideoMME (2.8). Removing r_{ts} causes smaller but consistent drops of 0.2–0.6 points. This pattern supports their complementary roles: r_{dm} provides the primary counterfactual signal, whereas r_{ts} acts as a temporal regularizer.

Effects of Frame Sampling Budget. We compare all models at matched budgets of 16, 32, and 64 frames (Figure 5). VIDEOMEND consistently achieves the highest mean accuracy, scoring 59.1%, 61.0%, and 61.1%, respectively. Its margin over Video-R1 remains 2.2–2.9 percentage points, showing that the advantage persists across the evaluated sampling budgets. Increasing the budget from 16 to 32 and from 32 to 64 frames yields gains of 1.9 and 0.1 points, respectively, indicating diminishing returns beyond 32 frames. maintaining an advantage across budgets.

4.4 TEMPORAL AND VISUAL FAITHFULNESS ANALYSIS

We compare SFT-only with VIDEOMEND to assess temporal faithfulness after DM-GRPO. On TempCompass, temporal hallucinations, defined as timestamps exceeding the video duration, decrease from 10.7% to 0.7% (Figure 6a). Mean temporal coverage increases from 89.4% to 99.1%, and the proportion of samples with over 80% coverage rises from 84.1% to 98.5% (b–c). Human evaluation of 500 Deltas per model further shows an increase in visual correctness from 63.4% to 92.2% (d). Annotators judge descriptions against the corresponding video intervals without access to generated or ground-truth answers (Appendix C.4). Together, these results support improved temporal faithfulness after DM-GRPO, with fewer timestamp violations, broader temporal coverage, and more visually grounded Delta descriptions.

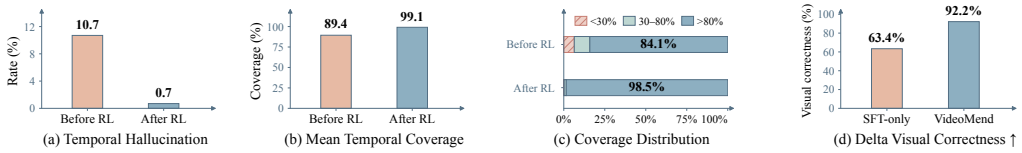


Figure 6: Temporal and visual faithfulness before and after DM-GRPO: (a) timestamp violations, (b) mean temporal coverage, (c) coverage distribution, and (d) Delta visual correctness. Before/after RL denotes SFT-only/VIDEOMEND.

5 CONCLUSIONS

We present **VIDEOMEND** to address temporal discontinuity and single-frame bias in video reasoning under sparse frame sampling. When only a few frames are observed, a model must infer how the scene evolves between them, yet it may instead rely on static visual cues or language priors. To make inter-frame dynamics explicit, **VIDEOMEND** introduces Ref-Delta CoT, which represents sampled frames as references and describes the changes between them as Deltas. We further use temporal direction and speed augmentation to support cold-start training. Building on this representation, DM-GRPO constructs counterfactual inputs by masking Deltas and uses reward signals to discourage predictions that remain unchanged when motion evidence is removed. This encourages the model to ground its reasoning in visual changes rather than static frames or language priors. Across multiple video benchmarks, **VIDEOMEND** consistently improves upon the base model, with particularly strong gains on temporal reasoning tasks. These results highlight the value of explicitly modeling inter-frame dynamics for video understanding. Further case analyses, limitations, and broader impacts are provided in Appendices E, F, and G.

AI USE STATEMENT

Generative AI tools were used to polish the language and improve the readability of the manuscript. As part of the research pipeline, Qwen3.5-122B-A10B generated Ref–Delta reasoning annotations and provided offline Judge labels for DM-GRPO, as described in Section 3.3 and Appendix A.4. Model-based filtering was also used to detect answer leakage in generated annotations (Appendix A.3). Quality filtering and human audits are documented in Appendices A.3 and B.2–B.4. The authors take full responsibility for the final manuscript, methodology, and reported results.

ETHICS STATEMENT

VideoMend studies temporal video reasoning using existing video datasets, constructed temporal augmentations, and model-generated supervision. The resulting models may inherit biases from the source data, backbone, and annotation model. Although quality filtering and human audits assess supervision reliability, they do not eliminate annotation errors or guarantee faithful reasoning. Ref–Delta traces are model-generated explanations and should not be treated as definitive evidence in high-stakes decisions. Applications involving videos of people should respect privacy and consent and account for potential misuse in surveillance or behavior monitoring. Further limitations and broader impacts are discussed in Appendices F and G.

REPRODUCIBILITY STATEMENT

Sections 3.1–3.3 describe data collection, temporal augmentation, and Ref–Delta annotation generation, while Section 3.4 specifies DM-GRPO. Section 4.1 reports the training configuration, hardware, benchmarks, and evaluation protocol. Appendices A.1–A.4 detail the data composition, prompt templates, quality filtering, and offline Delta-Mask data construction, respectively. Appendix B.1 presents training curves, and Appendices B.2–B.4 document human validation of Judge labels, retained SFT traces, and Delta visual correctness. Appendices C.1–C.3 provide matched-frame comparisons, broader generalization results, and inference-efficiency measurements, respectively.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Tieyuan Chen, Huabin Liu, Tianyao He, Yihang Chen, Chaofan Gan, Xiao Ma, Cheng Zhong, Yang Zhang, Yingxue Wang, Hui Lin, and Weiyao Lin. Mecd: Unlocking multi-event causal discovery in video reasoning. In *Advances in Neural Information Processing Systems*, 2024a.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziqi Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Lost in time: A new temporal benchmark for video llms. *arXiv preprint arXiv:2410.07752*, 2024.

Jisheng Dang, Jingze Wu, Teng Wang, Xuanhui Lin, Nannan Zhu, Hongbo Chen, Wei-Shi Zheng, Meng Wang, and Tat-Seng Chua. Reinforcing Video Reasoning with Focused Thinking. *arXiv preprint arXiv:2505.24718*, 2025.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Nature*, 645:633–638, 2025.

Sunqi Fan, Jiashuo Cui, Meng-Hao Guo, and Shuojin Yang. Tool-augmented spatiotemporal reasoning for streamlining video question answering task. *arXiv preprint arXiv:2512.10359*, 2025.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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 (CVPR)*, pp. 24108–24118, June 2025.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "Something Something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5842–5850, 2017.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, et al. M-llm based video frame selection for efficient video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13702–13712, 2025a.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025b.

Pengfei Hu, Meng Cao, Yingyao Wang, Yi Wang, Jiahua Dong, Jun Song, Yu Cheng, Bo Zheng, and Xiaodan Liang. Thinking with drafts: Speculative temporal reasoning for efficient long video understanding. *arXiv preprint arXiv:2512.00805*, 2025c.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025a.

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.

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*, 2025b.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pp. 323–340. Springer, 2024b.

Zongzhao Li, Zongyang Ma, Mingze Li, Songyou Li, Yu Rong, Tingyang Xu, Ziqi Zhang, Deli Zhao, and Wenbing Huang. Star-r1: Spatial transformation reasoning by reinforcing multimodal llms. *arXiv preprint arXiv:2505.15804*, 2025c.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26689–26699, 2024.

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024a.

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, 2024b.

Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. When thinking drifts: Evidential grounding for robust video reasoning. *arXiv preprint arXiv:2510.06077*, 2025.

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

Muhammad Maaz, Hanoona Rasheed, Fahad Shahbaz Khan, and Salman Khan. Video-r2: Reinforcing consistent and grounded reasoning in multimodal language models. *arXiv preprint arXiv:2511.23478*, 2025.

Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Hanoona Rasheed, Abdelrahman Shaker, Anqi Tang, Muhammad Maaz, Ming-Hsuan Yang, Salman Khan, and Fahad S. Khan. Videomathqa: Benchmarking mathematical reasoning via multimodal understanding in videos. *arXiv preprint arXiv:2506.05349*, 2025.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Sicheng Tao, Jungang Li, Yibo Yan, Junyan Zhang, Yubo Gao, Hanqian Li, Shuhang Xun, Yuxuan Fan, Hong Chen, Jianxiang He, and Xuming Hu. Moss-chatv: Reinforcement learning with process reasoning reward for video temporal reasoning. *arXiv preprint arXiv:2509.21113*, 2025.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22958–22967, October 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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.

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*, pp. 9777–9786, 2021.

Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. In *Advances in Neural Information Processing Systems*, 2025.

Ziang Yan, Yinan He, Xinhao Li, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1.5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8657–8670, June 2025.

Songyuan Yang, Weijiang Yu, Jilin Ma, Ziyu Liu, Guijian Tang, Wenjing Yang, Huibin Tan, and Nong Xiao. Reinforce to learn, elect to reason: A dual paradigm for video reasoning. *arXiv preprint arXiv:2604.04379*, 2026.

En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, et al. Unhackable temporal reward for scalable video mllms. In *International Conference on Learning Representations (ICLR)*, 2025.

Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. *arXiv preprint arXiv:2501.07888*, 2025.

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*, 2025a.

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. *Transactions on Machine Learning Research*, 2025b.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22056–22065, October 2025c.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research (TMLR)*, 2024.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8475–8489, June 2025.

APPENDIX CONTENTS

A Extended Related Work 15

B Data Construction and Annotation Details 15

B.1 VIDEOMEND-RefDelta-CoT-13K Distribution 15

B.2 Prompt Templates 15

B.3 VIDEOMEND-RefDelta-CoT-13K Quality Filtering 17

B.4 Offline DM-Data-8K Construction 18

C Training Diagnostics and Reliability Evaluation 19

C.1 Training Curves 20

C.2 Offline Judge Human Validation 20

C.3 Human Audit of Retained Ref–Delta SFT Traces 21

C.4 Independent Human Evaluation of Delta Visual Correctness 21

D Additional Experiments 21

D.1 Broader Generalization Benchmarks 23

D.2 Inference Efficiency 23

E Qualitative Case Analysis 24

E.1 TempCompass Case 25

E.2 MVBench Case 25

F Limitations and Future Work 25

G Broader Impacts 25

A Extended Related Work

Video Large Language Models. Recent years have witnessed substantial progress in Video Large Language Models. Video-ChatGPT Maaz et al. (2024) established an early architecture for aligning video representations with large language models. Subsequent methods such as VideoLLaMA 2 Cheng et al. (2024) and LLaVA-Video Zhang et al. (2024) advanced spatiotemporal modeling and video instruction tuning, while Qwen2.5-VL Bai et al. (2025), InternVL2.5 Chen et al. (2024b), and VideoLLaMA 3 Zhang et al. (2025a) further improved general-purpose video understanding.

Despite these advances, the computational cost of visual tokens requires most Video LLMs to process only a sparse subset of the continuous video. A related line of work therefore studies frame or keyframe selection Hu et al. (2025a); Zhang et al. (2025c); Hu et al. (2025c); Zhang et al. (2025b), aiming to retain more informative visual evidence under a limited sampling budget. Such methods primarily optimize which frames are provided to the model. VideoMend is complementary to this direction: given sampled frames, it focuses on how the reasoning process represents their temporal relationships by explicitly modeling inter-frame changes.

Temporal Video Understanding. Temporal reasoning has also been studied through benchmarks and specialized models targeting temporal perception, localization, and compositional understanding Liu et al. (2024b); Huang et al. (2024); Ren et al. (2024); Li et al. (2024a); Fu et al. (2025); Wu et al. (2024); Yang et al. (2025). These studies highlight persistent difficulties in recognizing motion, ordering events, and reasoning over long temporal contexts. Our work focuses on a complementary question: how the reasoning trace itself can explicitly represent temporal evidence distributed between sampled observations.

Reinforcement Learning for Video Reasoning. Inspired by RL-based reasoning in language models DeepSeek-AI et al. (2025); Wei et al. (2022), recent methods have introduced reinforcement learning into video reasoning. Video-R1 Feng et al. (2025) employs RLVR and temporal rewards; VideoRFT Wang et al. (2025a) introduces reasoning supervision and cross-modal consistency; and VideoChat-R1 Li et al. (2025b) and VideoChat-R1.5 Yan et al. (2025) explore spatiotemporal rewards and test-time scaling. STAR-R1 Li et al. (2025c), DeepVideo-R1 Park et al. (2025), and RLER Yang et al. (2026) further investigate spatial reasoning, difficulty-aware optimization, and evidence-aligned reasoning, respectively. Other recent studies likewise improve temporal reasoning and visual grounding through specialized reward designs Dang et al. (2025); Luo et al. (2025); Yu et al. (2025).

VideoMend differs primarily in the target of optimization. Rather than only rewarding answer correctness or general temporal/visual consistency, we explicitly structure the reasoning process into References and Deltas and construct counterfactual masked inputs to test whether individual Deltas are supported by their corresponding visual evidence. Thus, DM-GRPO directly optimizes the visual faithfulness of inter-frame reasoning.

B Data Construction and Annotation Details

This section provides additional details about the data and annotation pipeline used in VideoMend. We first describe the composition of VideoMend-RefDelta-CoT-13K and the prompts used to generate Ref–Delta reasoning traces. We then present the quality-control procedure applied to the generated annotations and the construction of the offline Delta-Mask dataset used by DM-GRPO.

B.1 VideoMend-RefDelta-CoT-13K Distribution

Table 3 summarizes the source distribution of VideoMend-RefDelta-CoT-13K. The dataset combines broad-coverage base video samples with targeted temporal augmentation data, resulting in 13,248 Ref–Delta CoT annotations in total.

B.2 Prompt Templates

This subsection collects the prompt templates used during annotation generation, training, and inference. We first present the answer-guided prompt used to synthesize Ref–Delta CoT supervision,

Table 3: Source distribution of VIDEOMEND-RefDelta-CoT-13K.

Category	Source	Count	Ratio
Base Dataset	ActivityNet	2,571	19.4%
	Charades	1,870	14.1%
	YouCook2	1,538	11.6%
	CLEVRER	1,530	11.6%
	NextQA	1,200	9.1%
	YouTube Videos	753	5.7%
	STAR	388	2.9%
	PerceptionTest	362	2.7%
	Ego4D	180	1.4%
	Subtotal	*10,392*	*78.4%*
Augmentation Dataset	Temporal Direction	1,340	10.1%
	Temporal Speed	1,516	11.4%
	Subtotal	*2,856*	*21.6%*
Total		**13,248**	**100%**

Prompt Template for Ref-Delta CoT Generation

```
System Prompt:
You are an expert video analyst. Analyze the "Semantic Flow" of a video by tracking dynamic
changes across time. Ignore static backgrounds unless they interact with the main subject.

CRITICAL: Video frames have HH:MM:SS timestamps watermarked in the top-left corner. Use ONLY exact
timestamps you observe — never estimate or fabricate. {fps_constraint}
Output format:
1. Reference Frame: "Reference frame at <time>HH:MM:SS</time>: [brief description]."
2. Delta: "From <time_range>HH:MM:SS - HH:MM:SS</time_range>, <delta>[what changed]</delta>."
3. Reasoning Summary: "<reasoning_summary>[connect deltas to answer]</reasoning_summary>"
4. Answer: "<answer>concise answer</answer>"

Example:
Reference frame at <time>00:00:00</time>: a person holds a red cup, white bowl to the left.
From <time_range>00:00:00 - 00:00:03</time_range>, <delta>the person tilts the cup and pours
liquid into the bowl</delta>.
From <time_range>00:00:03 - 00:00:05</time_range>, <delta>the person pushes the filled bowl
forward</delta>.
<reasoning_summary>The person poured liquid from cup to bowl, then pushed it
away.</reasoning_summary>
<answer>The person poured liquid into the bowl.</answer>

Cover the entire video timeline. Each delta must describe a different change — never repeat
previous observations. If an action repeats, summarize it in one delta instead of describing each
repetition. Keep your analysis concise.

Question: {question}
The correct answer is given below. You MUST logically arrive at this conclusion
by connecting the temporal deltas.
{answer}

Users:
<video>[base64-encoded video]</video>
Please analyze the video and answer the question.
```

Figure 7: Answer-guided prompt for Ref–Delta CoT annotation generation. The ground-truth answer is provided to guide faithful reasoning-chain construction.

followed by the prompt used during DM-GRPO training and evaluation without access to ground-truth answers. Finally, we provide the augmentation-specific hint prefixes used only during annotation generation.

B.2.1 Answer-Guided CoT Generation Prompt

Figure 7 shows the prompt used to generate Ref–Delta CoT annotations for SFT training data. The ground-truth answer is provided explicitly, and the model is instructed to construct a reasoning chain that faithfully arrives at that conclusion based on the video evidence. The system prompt additionally embeds an FPS constraint computed from the actual sampling rate.

B.2.2 Training and Inference Prompt

Figure 8 shows the prompt used during both DM-GRPO training and evaluation. Unlike the annotation-generation prompt, no ground-truth answer is provided. The model reasons over the video autonomously and produces a Ref–Delta CoT trace. The same prompt and output schema are used at evaluation time, with answers extracted from the `<answer>` field.

Dimension-Specific Hint Prefixes for Augmentation Data

```
Speed: NOTE: This is a speed-related question. Your analysis should reflect how fast or
slow the action is happening in the video.

Direction: NOTE: This is a direction-related question. Your analysis should reflect
which direction the subject is moving in the video.

Temporal: NOTE: This question asks about temporal order. Your analysis should reflect
whether the sequence of events flows naturally forward or appears reversed.
```

Figure 8: Prompt template used during DM-GRPO training and evaluation. The model reasons autonomously and outputs a Ref–Delta CoT trace without access to the ground-truth answer.

B.2.3 Augmentation Hint Prefixes

Figure 9 summarizes the dimension-specific hint prefixes appended during CoT annotation generation for the two augmentation subsets. The direction prefix orients the model toward describing motion direction in each Delta, while the speed prefix directs attention to inter-frame transition rates. These prefixes are used only during annotation generation and are absent at training and inference time.

B.3 VideoMend-RefDelta-CoT-13K Quality Filtering

We apply a three-stage quality filter consisting of structural rules, semantic rules, and a judge-based leakage check. The first two stages are rule-based and comprise 15 rules in total, while the final stage uses a judge model to detect subtle answer leakage that lexical rules may miss. Together, these stages filter raw annotations produced by the CoT generation pipeline before they are used for SFT.

Stage 1: Structural validation. The first stage enforces the structural validity of Ref–Delta CoT annotations and consists of eleven rules (R1–R11). A sample is rejected if it contains upstream error markers, if the `<think>` block is truncated or its body is shorter than 200 characters, or if any required structural element is missing. These required elements include `<time>`/`<time_range>` timestamps, `<delta>` tags, `<reasoning_summary>` tags, `<answer>` tags, and the "Reference frame at" anchor phrase. Samples whose corresponding video files are missing from disk are also discarded in this stage. Rejection is applied at the first triggered rule, and subsequent rules are not evaluated.

Stage 2: Semantic rule-based filtering. The second stage applies four semantic rules (R12–R15) to remove invalid or unreliable annotations. Meta-language detection (R12) rejects responses containing phrases that indicate model refusal or self-disclosure, such as "as an AI language model" or "I cannot process". Answer-leakage detection (R13) rejects responses whose reasoning chain explicitly

Prompt Template for Training and Inference

System Prompt:
A conversation between User and Assistant. The user asks a question about a video, and the Assistant solves it. The assistant first analyzes the video by establishing reference frames with <time> tags and tracking temporal changes with <time_range> and <delta> tags, then summarizes the reasoning in <reasoning_summary> tags. The analysis and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> analysis here </think><answer> answer here </answer>

Users:
<video>  </video>
{Question} {OutputTemplate}

OutputTemplate:
"Multiple Choice": Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.
"Numerical": Please provide the numerical value (e.g., 42 or 3.14) within the <answer> </answer> tags.
"OCR": Please transcribe text from the image/video clearly and provide your text answer within the <answer> </answer> tags.
"Free-form": Please provide your text answer within the <answer> </answer> tags.

Figure 9: Dimension-specific hint prefixes appended during augmentation annotation generation. Direction prefixes target motion directionality, while speed prefixes target inter-frame transition rates.

refers to the ground-truth answer through patterns such as "ground truth" or "answer key". For multiple-choice samples, which are identified by the presence of option lines in the question and a parseable A–E ground-truth label, R14 rejects samples whose predicted answer cannot be extracted, and R15 rejects samples where the extracted prediction does not match the ground truth. Free-form samples are exempt from answer verification to avoid false positives caused by annotation diversity. All rejected samples are saved with a `_reject_reason` field for error analysis.

Stage 3: Judge-based leakage detection. After the two rule-based stages, each retained sample is passed through an additional judge-based filtering step. The judge receives the reasoning chain together with the ground-truth answer and determines whether the reasoning is genuinely derived from video content or is implicitly constructed around the provided answer. Samples flagged by the judge as leaked are discarded. This final stage complements the lexical leakage detection in R13 with a semantic-level leakage check.

B.4 Offline DM-Data-8K Construction

This subsection describes the three-stage pipeline used to construct the offline mask patterns and Judge labels required by DM-GRPO (Section 3.4). Figure 10 summarizes the resulting DM-Data statistics.

Stage 1: Ref–Delta CoT generation. Following the procedure in Section 3.3, we first generate Ref–Delta CoT annotations for each training sample. After quality filtering, **8,128** annotations containing structured `<time_range>` tags are retained as input to the subsequent mask pattern construction.

Stage 2: Delta-to-frame mapping. For each annotation, we extract the ordered sequence of `<time_range>` tags, obtaining Delta intervals $\{[s_d, e_d]\}_{d=1}^{D}$ in seconds. Frame timestamps are computed by replicating the `smart_nframes` sampling schedule used at training time:

$$t_k = \frac{k-1}{N-1} \cdot \frac{F_{\text{total}} - 1}{\text{fps}}, \quad k = 1, \ldots, N, \tag{9}$$

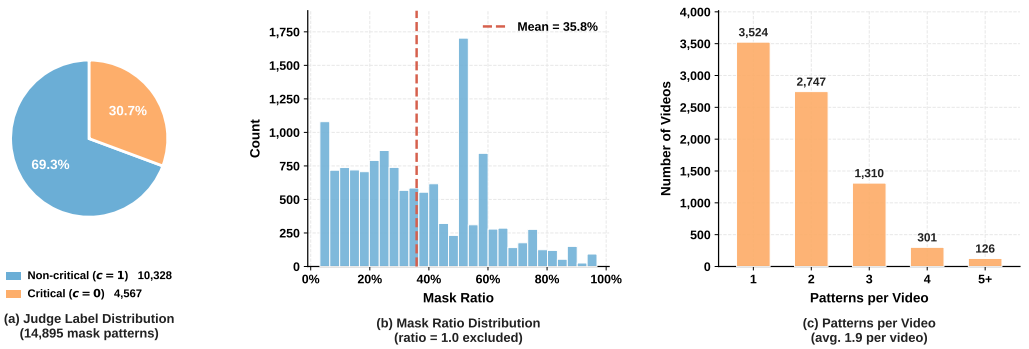


Figure 10: Statistics of DM-Data (8,008 video–question pairs, 14,895 mask patterns). **(a)** 69.3% of patterns are labeled non-critical ($c = 1$). **(b)** Mask ratio distribution (mean $= 35.8\%$). **(c)** Most videos yield one or two mask patterns (avg. 1.9).

where N is the number of sampled frames, clamped to $[4, N_{\max}]$ and rounded to the nearest even number. The frame index set for Delta d is then

$$\mathcal{I}_d = \{k : \lfloor t_k \rfloor \in [\lfloor s_d \rfloor, \lfloor e_d \rfloor]\}\,. \tag{10}$$

Strict replication of the sampling schedule is necessary: any mismatch between construction-time and training-time frame indices would cause mask patterns to silently cover the wrong frames during policy optimization. Each $\mathcal{I}_d$ with $N - |\mathcal{I}_d| \geq 4$ defines a valid mask pattern; patterns below this threshold are discarded to ensure sufficient visible context. If all patterns for a video fail this check, the one with the most remaining frames is retained as a fallback.

This procedure yields **14,895** valid mask patterns across **8,110** videos, with an average of 1.9 patterns per video and an average mask ratio of 40.3%.

Stage 3: Judge pre-annotation. For each mask pattern $(\mathcal{I}_d, q)$, we construct the masked input $\tilde{F}$ by zeroing the pixels of $\{f_k\}_{k \in \mathcal{I}_d}$ while preserving their temporal positions, so the sequence length and 3D RoPE positional encodings remain unchanged. To prevent API-side re-sampling, each frame is encoded independently as a base64 JPEG (quality 85) and submitted as a multi-image sequence rather than a video file.

Qwen3.5-122B-A10B is then queried without chain-of-thought (`max_tokens=128`) for a direct single-letter answer. The Judge label is defined as

$$c = \mathbb{I}[\text{Judge answers } q \text{ correctly on } \tilde{F}]. \tag{11}$$

A correct answer indicates a non-critical interval ($c = 1$); otherwise the interval is critical ($c = 0$). Inference runs with 8 concurrent threads and completes in approximately 2 hours for all 14,895 patterns. The final dataset contains **8,008** video–question pairs, with 69.3% of patterns labeled $c = 1$, providing a substantial proportion of counterfactually informative training signal for DM-GRPO.

Implementation note. $N_{\max}$ must be set identically in Stage 2 and during DM-GRPO training; any mismatch causes silent frame-index misalignment that corrupts the mask signal. We fix $N_{\max} = 16$ throughout.

C Training Diagnostics and Reliability Evaluation

This section examines the optimization dynamics of DM-GRPO and provides human validation of the supervision signals used during training. We first report the evolution of the individual reward components and then evaluate the reliability of offline Judge labels, retained SFT traces, and generated Delta descriptions.

C.1 Training Curves

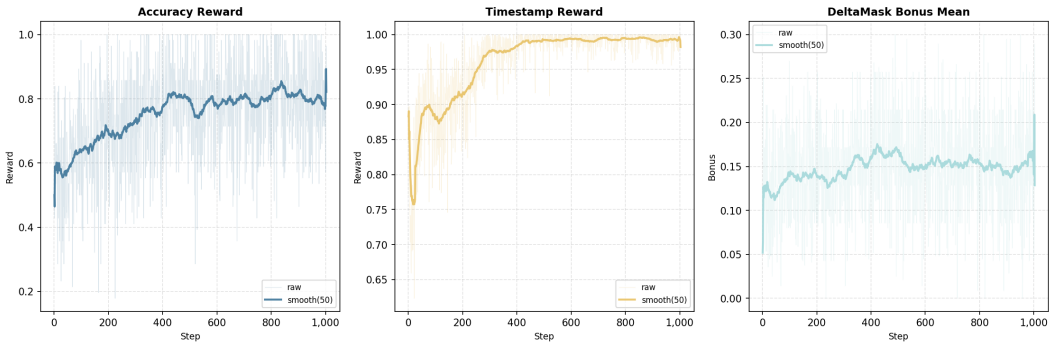


Figure 11: Accuracy, timestamp, and DeltaMask reward dynamics during DM-GRPO training.

Figure 11 shows the reward dynamics during DM-GRPO training, including accuracy reward, timestamp reward, and DeltaMask bonus. The accuracy reward steadily increases during training, with a rapid rise at the beginning, a middle-stage plateau, and a further improvement near the end. Despite the variance in the raw curve from group-based reward estimation, the smoothed trend shows consistent improvement in task-level performance.

The timestamp reward converges quickly after a brief early fluctuation and remains near saturation, indicating that the model rapidly learns valid and temporally coherent Ref–Delta structures and that timestamp supervision mainly stabilizes the reasoning format. The DeltaMask bonus increases rapidly at the beginning and then grows more gradually. Importantly, the bonus does not collapse during training, indicating that the counterfactual signal continues to encourage visually grounded Delta reasoning throughout optimization.

Overall, the three rewards play complementary roles: timestamp reward stabilizes the Ref–Delta structure, accuracy reward drives task-level improvement, and DeltaMask bonus promotes visual grounding of temporal Deltas.

C.2 Offline Judge Human Validation

We manually audit 800 Delta-Mask patterns, comprising 400 labeled non-critical ($c = 1$) and 400 labeled critical ($c = 0$) by the offline Judge.

Validation criterion. Human annotators assess whether the visual evidence remaining after masking is sufficient to answer the question. A pattern is judged non-critical if the question remains answerable from that evidence, even when the masked interval contains useful information; otherwise, it is judged critical. The assessment concerns evidence sufficiency rather than whether an answer can be guessed from general knowledge or language priors.

Results. We compare the human judgments with the offline Judge labels. As shown in Table 4, annotators disagree with 16 of 400 non-critical labels and 10 of 400 critical labels, yielding 96.75% agreement across the balanced audit sample. This supports the reliability of the Judge labels under the evaluated criterion; the reported agreement is specific to the balanced sample, rather than weighted by the full dataset's label distribution.

Table 4: Human agreement with offline Judge labels on a class-balanced sample of 800 Delta-Mask patterns.

Judge label	Audited	Disagree	Disagree rate	Agreement
Non-critical ($c = 1$)	400	16	4.00%	96.00%
Critical ($c = 0$)	400	10	2.50%	97.50%
Overall (balanced)	800	26	3.25%	**96.75%**

C.3 Human Audit of Retained Ref–Delta SFT Traces

We manually inspect 1,000 Ref–Delta SFT traces retained after quality filtering (Appendix B.3) to assess the quality of teacher-generated supervision.

Audit criteria. Each trace is assessed for (i) visual consistency of Reference descriptions; (ii) support for Delta descriptions within the stated intervals; (iii) correspondence between timestamps and described events; and (iv) consistency of the reasoning summary and final answer with the visual evidence. A trace is marked correct only if all criteria are satisfied. Unlike the Delta-level evaluation in Appendix C.4, this audit assesses the complete supervision trace.

Results. Of the 1,000 audited traces, 986 are judged correct and 14 incorrect (Table 5), yielding 98.6% correctness on the audited sample. This result supports the quality of the retained SFT supervision.

Table 5: Human audit of 1,000 retained Ref–Delta SFT traces.

Outcome	Count	Rate
Correct	986	**98.6%**
Incorrect	14	1.4%

C.4 Independent Human Evaluation of Delta Visual Correctness

We manually evaluate 1,000 generated Deltas, with 500 from SFT-only and 500 from VideoMend. Annotators inspect each Delta alongside its corresponding video interval, without access to generated or ground-truth answers.

Evaluation criteria. A Delta is considered visually correct if (i) the described entities and actions are supported by the video; (ii) the stated change, including direction, order, or speed when specified, agrees with the observed transition; and (iii) the description introduces no unsupported events or details. The assessment concerns the visual support for the described change, rather than its relevance to the final answer or the correctness of that answer.

Results. As shown in Table 6, visual correctness increases from 63.4% for SFT-only to 92.2% for VideoMend, a gain of 28.8 percentage points. This comparison provides direct human evidence that the Delta descriptions generated by VideoMend are more consistent with the corresponding visual evidence.

Table 6: Human evaluation of visual correctness for 500 generated Deltas per model.

Model	Evaluated	Correct	Incorrect	Visual correctness
SFT-only	500	317	183	63.40%
VideoMend	500	461	39	**92.20%**

D Additional Experiments

This section reports additional experimental results beyond those included in the main paper. For the six main benchmarks, we provide matched-frame comparisons against Video-R1 under 16-, 32-, and 64-frame settings and include a published 7B reasoning baseline where directly comparable. We also report a controlled ablation of the explicit Ref–Delta format, additional temporal-reasoning benchmarks, broader generalization benchmarks, and inference-efficiency measurements.

D.0.1 AoTBench

We compare against published AoTBench baselines from ArrowRL Xue et al. (2025). The reported overall score is the sample-count-weighted accuracy over the five sub-evaluations.

Table 7: AoTBench results with published comparisons. Published model results are reproduced from Xue et al. (2025). "Overall" is recomputed as sample-count-weighted accuracy over the five sub-evaluations (144 ReverseFilm, 500 UCF101, 1,992 T2V, 1,992 V2T, and 1,800 AoT-VQA examples). Bold marks the best value among the listed models.

Method	ReverseFilm	UCF101	T2V	V2T	AoT-VQA	Overall
Random chance	50.0	50.0	50.0	50.0	39.3	47.00
GPT-4o Xue et al. (2025)	52.8	54.0	56.5	69.5	**67.8**	**63.42**
Gemini-1.5-Pro Xue et al. (2025)	51.4	52.8	60.4	58.9	57.3	58.27
Aria Xue et al. (2025)	50.0	51.6	56.4	57.5	55.9	56.08
MiniCPM-V-2.6 Xue et al. (2025)	50.0	51.2	54.8	61.3	46.8	54.19
InternLM-XC-2.5 Xue et al. (2025)	50.0	51.6	52.0	53.6	45.8	50.68
LLaVA-Video-7B Xue et al. (2025)	50.0	51.6	57.2	63.1	46.7	55.49
Video-R1-7B Xue et al. (2025)	50.0	51.6	56.3	62.5	46.7	55.03
LLaVA-OV-7B Xue et al. (2025)	50.0	51.6	56.3	62.4	46.2	54.86
LLaVA-OV-7B + ArrowRL Xue et al. (2025)	54.2	57.4	**57.9**	66.1	53.2	59.00
Qwen2-VL-7B Xue et al. (2025)	50.0	51.6	56.3	62.3	44.3	54.29
Qwen2-VL-7B + ArrowRL Xue et al. (2025)	**69.1**	**72.6**	57.1	68.8	51.1	60.52
Qwen2.5-VL-7B Xue et al. (2025)	50.0	51.6	53.4	66.6	49.6	56.21
Qwen2.5-VL-7B + ArrowRL Xue et al. (2025)	51.4	54.8	55.6	69.6	58.8	60.68
VIDEOMEND	50.0	51.6	57.63	**70.58**	55.28	60.35

Table 8: TVBench per-task results with published video-input baselines. Original benchmark results are from Cores et al. (2024); Tarsier2-7B results are from Yuan et al. (2025). Macro is the unweighted mean of the ten task accuracies. Tarsier2-7B's source reports 54.7% as an item-weighted overall score; its 52.6 macro mean is recomputed from the displayed task scores to match this table's aggregation. Bold marks the best non-human result in each column.

Method	AC	OC	AS	OS	ST	AL	AA	UA	ES	MD	Macro
Published TVBench baselines (video input)											
VideoLLaVA Cores et al. (2024)	33.8	27.0	54.7	33.8	57.8	23.1	45.7	24.1	22.5	23.3	34.6
VideoChat2 Cores et al. (2024)	25.9	27.0	56.8	32.9	56.2	23.1	48.1	32.9	24.5	22.4	35.0
ST-LLM Cores et al. (2024)	25.0	35.1	51.7	36.0	54.4	31.0	45.6	34.2	24.0	20.3	35.7
GPT-4o Cores et al. (2024)	26.1	21.3	59.3	33.2	52.4	25.0	78.4	41.7	31.0	30.6	39.9
PLLaVA-7B Cores et al. (2024)	32.1	25.7	55.6	33.3	52.4	23.8	53.1	30.5	20.5	21.6	34.9
PLLaVA-13B Cores et al. (2024)	37.3	24.3	61.8	33.3	55.1	28.1	47.8	39.0	19.5	17.2	36.3
PLLaVA-34B Cores et al. (2024)	27.6	32.4	67.0	35.6	77.8	44.4	58.8	32.9	27.0	19.4	42.3
mPLUG-Owl3 Cores et al. (2024)	27.4	32.4	69.8	37.3	76.8	43.1	56.9	34.2	18.0	26.3	42.2
VideoLLaMA2.1 Cores et al. (2024)	25.4	30.4	71.2	29.8	76.6	46.9	58.1	36.8	23.5	22.4	42.1
VideoLLaMA2-7B Cores et al. (2024)	30.6	37.8	69.6	33.8	68.1	40.6	56.9	43.9	24.5	22.8	42.9
VideoLLaMA2-72B Cores et al. (2024)	26.7	46.6	73.5	40.4	81.6	53.1	76.6	36.6	19.5	29.7	48.4
VideoGPT+ Cores et al. (2024)	30.6	52.0	61.3	36.9	69.2	40.0	53.1	29.7	16.0	28.5	41.7
Gemini-1.5-Pro Cores et al. (2024)	33.5	22.3	71.5	34.5	82.6	51.6	77.8	43.9	25.0	33.6	47.6
Qwen2-VL-7B Cores et al. (2024)	27.1	61.5	63.8	38.7	75.7	41.3	63.1	22.0	22.5	22.4	43.8
Qwen2-VL-72B Cores et al. (2024)	32.6	**73.0**	68.6	31.1	82.2	45.0	75.3	37.8	19.5	62.1	52.7
LLaVA-Video-7B Cores et al. (2024)	32.6	23.6	78.5	32.9	81.6	58.8	71.9	29.3	22.5	24.1	45.6
LLaVA-Video-72B Cores et al. (2024)	38.6	27.0	80.3	33.3	85.9	**65.6**	66.6	35.4	25.5	41.8	50.0
IXC-2.5-7B Cores et al. (2024)	32.1	50.7	77.1	37.3	78.4	46.9	60.0	41.7	21.0	**70.3**	51.6
Aria Cores et al. (2024)	**58.4**	60.1	75.1	**42.2**	81.6	43.1	70.3	32.9	21.0	25.0	51.0
Tarsier-7B Cores et al. (2024)	22.9	58.8	73.2	35.6	64.9	46.9	75.6	40.2	25.5	25.0	46.9
Tarsier-34B Cores et al. (2024)	31.7	64.2	**81.7**	32.1	77.8	58.1	84.4	**52.4**	24.5	48.3	55.5
Additional published reference and our evaluations											
Human baseline Cores et al. (2024)	100.0	94.9	100.0	90.6	90.0	96.0	100.0	86.0	90.0	100.0	94.8
Tarsier2-7B Yuan et al. (2025)	40.5	62.8	72.3	31.6	**88.1**	37.5	**94.1**	41.5	24.5	33.2	52.6
Qwen2.5-VL-7B (our eval)	36.2	58.8	75.1	33.8	78.4	59.4	72.1	48.8	50.5	30.6	54.4
VIDEOMEND	36.8	54.1	73.7	37.8	84.9	60.0	78.2	42.7	**53.5**	41.8	**56.3**

D.0.2 TVBENCH

TVBench contains ten temporal tasks Cores et al. (2024). We add the public Tarsier2-7B per-task results Yuan et al. (2025); ArrowRL's published overall score on TVBench is also shown for reference Xue et al. (2025).

D.1 Broader Generalization Benchmarks

D.1.1 NExT-QA

We report causal, temporal, descriptive, and overall accuracies on the NExT-QA multiple-choice test set Xiao et al. (2021). Published test-set baselines are taken from STAR Fan et al. (2025).

Table 9: MECD results with all comparison methods from the original benchmark paper Chen et al. (2024a). Ave. SHD is lower-is-better; accuracy is higher-is-better. Dashes indicate that the metric was not reported for our evaluation.

Paradigm	Method	Ave. SHD	Accuracy (%)
Few-shot LLMs			
	Gemini-1.5-Pro	4.91	59.3
	GPT-4-0613	4.92	59.6
Video-language model baselines			
	MiniGPT4-video	5.16	56.8
	MiniGPT-4	5.14	57.5
	Video-LLaMA	5.10	60.6
	VideoChat2	4.89	60.7
	VideoLLaVA	4.85	62.5
	GPT-4o	4.69	65.5
Fine-tuned multimodal methods			
	VAR	4.96	57.3
	VideoBERT	4.95	60.9
	CLIP (ViT-L/14)	4.77	62.9
	VideoChat2 (fine-tuned)	4.77	66.9
	VideoLLaVA (fine-tuned)	4.73	67.1
	VGCM‡ (without causal inference)	4.51	67.0
	VGCM	**4.19**	**71.2**
Additional evaluations			
	Qwen2.5-VL-7B (our eval)	–	63.3
	VideoMend (ours)	–	68.1
Human	Deductive reasoning	2.05	87.2

Original-paper validation-set baselines. The original NExT-QA benchmark paper also reports validation-set results for classic VideoQA systems using GloVe and fine-tuned BERT question representations. We report these separately from the test-set table above because the split differs.

Table 10: Original NExT-QA validation-set baselines reported in the benchmark release **?**. Acc@C, Acc@T, and Acc@D are accuracies on causal, temporal, and descriptive questions; Acc is overall accuracy.

Method	GloVe				BERT-FT			
	C	T	D	All	C	T	D	All
BlindQA	26.89	30.83	32.60	30.60	42.62	45.53	43.89	43.76
EVQA	28.69	31.27	41.44	31.51	42.64	46.34	45.82	44.24
STVQA	36.25	36.29	55.21	39.21	44.76	49.26	55.86	47.94
CoMem	35.10	37.28	50.45	38.19	45.22	49.07	55.34	48.04
HME	37.97	36.91	51.87	39.79	46.18	48.20	58.30	48.72
HCRN	39.09	40.01	49.16	40.95	45.91	49.26	53.67	48.20
HGA	35.71	38.40	55.60	39.67	46.26	50.74	59.33	49.74
Human	87.61	88.56	90.40	88.38	87.61	88.56	90.40	88.38

D.1.2 MECD

We compare against the baselines reported in the original MECD paper Chen et al. (2024a). All values below use top-1 causal-relation accuracy.

D.2 Inference Efficiency

We measure both generation-only and end-to-end throughput under the same setup: 8× NVIDIA A800 GPUs, 32 input frames, greedy decoding, and batch size 1.

Table 11: MECD accuracy with published baselines.

Method	Accuracy (%)
Gemini-1.5-Pro Chen et al. (2024a)	59.3
GPT-4-0613 Chen et al. (2024a)	59.6
VideoChat2 Chen et al. (2024a)	60.7
VideoLLaVA Chen et al. (2024a)	62.5
Qwen2.5-VL-7B (our eval)	63.3
GPT-4o Chen et al. (2024a)	65.5
VideoLLaVA (fine-tuned) Chen et al. (2024a)	67.1
VIDEOMEND	68.1
VGCM Chen et al. (2024a)	**71.2**

Table 12: Inference throughput and latency.

Model	Gen. throughput (samples/s)	Gen. latency (s/video)	E2E throughput (samples/s)	E2E latency (s/video)
Qwen2.5-VL-7B	4.680	0.214	3.171	0.315
VIDEOMEND	3.040	0.329	2.323	0.431

Under this setup, explicit Ref–Delta decoding increases end-to-end latency from 0.315 to 0.431 seconds per video. This overhead reflects the additional generation required to produce structured temporal reasoning traces.



VideoMend:<think>
Reference frame at <time>00:00:00.0</time>: A pile of roasted coffee beans lies scattered on a dark textured surface. The beans appear stationary, with no visible object-level motion.
From <time_range>00:00:00.0 - 00:00:00.5</time_range>, <delta>the camera begins a slight rightward movement while keeping the coffee beans in focus. The beans remain clustered, with no individual bean showing noticeable motion</delta>.
From <time_range>00:00:01.1 - 00:00:02.2</time_range>, <delta>the camera continues a slow horizontal pan to the right, revealing more scattered beans while their relative arrangement remains largely unchanged</delta>.
From <time_range>00:00:02.8 - 00:00:03.9</time_range>, <delta>the rightward camera motion becomes clearer as the viewpoint sweeps across the bean pile. The apparent shift of the beans is caused by the changing camera perspective rather than independent bean movement</delta>.
From <time_range>00:00:04.4 - 00:00:05.0</time_range>, <delta>a diagonal pattern of beans appears to slide downward-right in the frame, but this is best explained as camera-induced apparent motion rather than natural rolling or tumbling</delta>.
From <time_range>00:00:05.5 - 00:00:10.5</time_range>, <delta>the camera maintains a steady rightward pan, continuously changing the visible portion of the bean pile while the beans themselves remain essentially stationary relative to one another</delta>.
Reasoning Summary: The video shows a stationary pile of coffee beans under continuous camera panning. Although the beans appear to shift across the frame, the displacement is caused by the changing viewpoint rather than independent motion of the beans.
</think>
<answer>B</answer>

Figure 12: TempCompass case. VIDEOMEND distinguishes a stationary pile of coffee beans from apparent motion caused by camera panning through Ref–Delta reasoning.

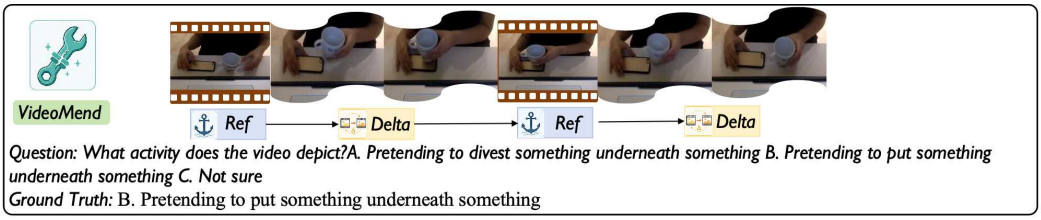
E QUALITATIVE CASE ANALYSIS

We provide qualitative examples from TempCompass and MVBench to illustrate how VIDEOMEND reasons with Ref–Delta CoT. Rather than relying on isolated frames, the model anchors static visual states with References and explains inter-frame changes with Deltas. This allows the reasoning trace to make explicit which temporal transitions support the final prediction.

E.1 TempCompass Case

In the TempCompass example shown in Figure 12, the apparent displacement of the coffee beans could be mistaken for object motion when individual frames are viewed in isolation. By comparing the visual state across adjacent observations, the Ref–Delta trace attributes the apparent change to camera movement rather than motion of the object itself.

E.2 MVBench Case



VideoMend:<think>
Reference frame at <time>00:00:00</time>: A person is seated at a table, with an iPhone near the left hand and a white mug near the right hand. Both objects are resting on the table surface.
From <time_range>00:00:00 - 00:00:02</time_range>, <delta>the person gently places the mug down on the table close to the phone, positioning it as if preparing to move it in relation to the phone</delta>.
Reference frame at <time>00:00:02</time>: The mug remains beside the phone on the table, and the person's hands are paused near the objects. The mug has not been placed underneath the phone.
From <time_range>00:00:02 - 00:00:03</time_range>, <delta>the person lifts the mug slightly and lowers it again near the phone, mimicking the motion of putting the mug underneath another object without actually sliding it under the phone</delta>.
From <time_range>00:00:03 - 00:00:04</time_range>, <delta>the person's hands return to their original position, and the mug remains on the table rather than being truly placed underneath the phone</delta>.
Reasoning Summary: The person moves the mug near the phone in a way that resembles putting one object underneath another, **but the mug is never actually placed under the phone.** This supports the action category of pretending to put something underneath something.
</think>
<answer>B. Pretending to put something underneath something</answer>

Figure 13: MVBench case. VideoMend analyzes the mug movement across adjacent frames and identifies the action as pretending to put something underneath another object.

In the MVBench example shown in Figure 13, the answer depends on a fine-grained temporal action rather than a single static state. The model tracks the mug movement across adjacent frames and uses the corresponding Deltas to identify the intended action. Together, the two examples illustrate how References anchor observed states while Deltas capture the changes required for temporal inference.

F Limitations and Future Work

Bounded by sampled evidence. VideoMend is designed to mend temporal discontinuity under the common sparse-frame sampling paradigm rather than replacing sampling itself. Therefore, its effectiveness is still bounded by the visual evidence preserved in sampled frames, especially when key events are extremely brief or entirely skipped. This reflects a broader computational constraint of current video MLLMs, and future work may combine Ref–Delta reasoning with adaptive or denser frame selection.

Toward richer temporal evidence. First, Ref–Delta reasoning can be integrated with adaptive frame selection, enabling the model to allocate more visual budget to temporally ambiguous or rapidly changing segments. Second, Delta grounding can be further enhanced with richer temporal signals, such as motion cues, optical flow, audio, or subtitles, to support more comprehensive video understanding beyond sampled RGB frames.

G Broader Impacts

Toward more faithful video reasoning. VideoMend encourages models to ground answers in inter-frame changes, improving the reliability and interpretability of video understanding for applications such as assistive analysis, robotics, education, and video retrieval.