

EventLens: Event-Structure Reinforcement Learning for Video Understanding

Guowen Zhang¹ Boshen Xu¹ Zihao Yue¹ Ziheng Wang¹
Xiaokun Liu^{2*} Xin Tao² Wenyu Qin² Pengfei Wan² Qin Jin^{1†}

¹AIM3 Lab, Renmin University of China ²Kling AI
Project Page: <https://lostgreen.github.io/EventLens/>

Abstract

Recent multimodal large language models (MLLMs) achieve strong results on video question answering, yet often fail to recover the temporal structure that makes a video coherent. They may merge neighboring events, miss single-event progression, or infer implausible relations across events. We study structured temporal understanding as the ability to recover event structure from visual evidence, including event boundaries, single-event progression, and multi-event relations. We argue that this gap arises from a mismatch between current post-training objectives and temporal reasoning, where caption-style supervision allows models to rely on language shortcuts rather than visual temporal grounding. To address this, we propose **EventLens**, a vision-centric reinforcement learning framework that learns video understanding through *event-structure recovery*. Instead of reconstructing captions, EventLens trains models to recover temporal structure under controlled perturbations, such as altered temporal granularity, reversed progression, and shuffled event order. We instantiate this framework with a three-level temporal ontology and derive three verifiable task families: event segmentation, progression discrimination, and multi-event relation reasoning. These tasks admit deterministic rewards, enabling task-specialized GRPO training without human preference labels or learned reward models. We further introduce multi-teacher on-policy distillation (MT-OPD) to consolidate specialized policies into a unified model. Experiments on Qwen3-VL backbones show consistent improvements on temporal grounding and event-sensitive reasoning benchmarks, while preserving general video-QA performance. Learning curves further demonstrate that EventLens achieves comparable downstream transfer with significantly reduced training cost compared to direct mixed RL optimization.

1 Introduction

Modern multimodal large language models (MLLMs) have achieved strong performance on a wide range of video question answering benchmarks [56, 45, 18]. However, this success largely reflects the ability to recognize what is visible, rather than to understand how events unfold over time. When tasks require explicit temporal structure, models often fail to separate events, track their progression, or reason about relations across them. As a result, models that appear capable on coarse-grained benchmarks can break down when deeper temporal understanding is required.

We study this limitation as *structured temporal understanding*: the ability to recover event boundaries, single-event progression, and multi-event relations from visual evidence. This capability underlies tasks such as temporal localization, progression analysis, and multi-event reasoning [28, 3]. Unlike

† Corresponding Author; * Industry Project Lead.

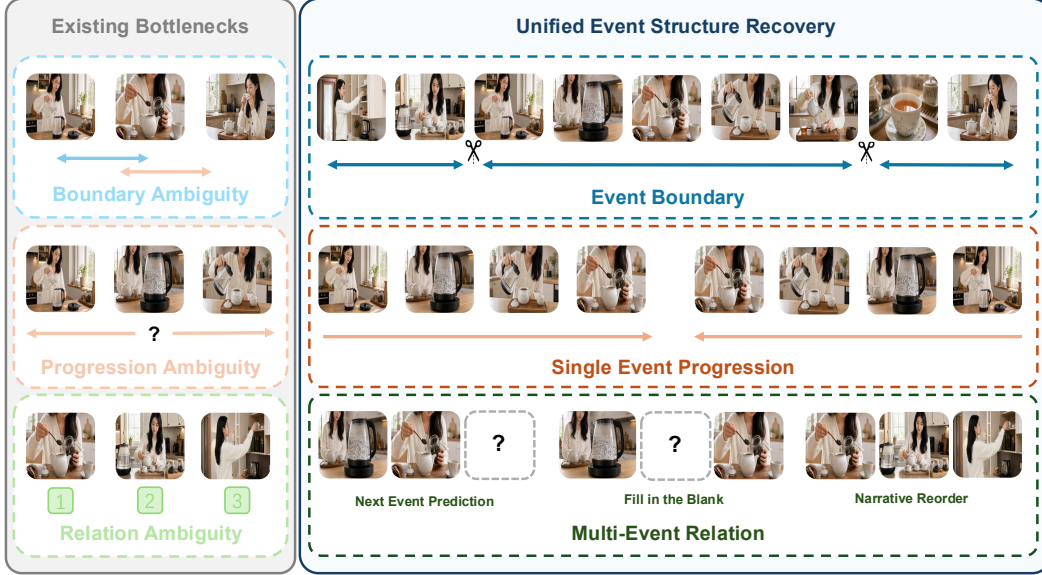


Figure 1: Motivation and task view of unified event-structure recovery. Existing video MLLMs often suffer from boundary ambiguity, progression ambiguity, and relation ambiguity when they must recover explicit temporal structure. We address these bottlenecks with three event-structure recovery tasks: event-boundary recovery, single-event progression discrimination, and multi-event relation reasoning through next-event prediction, fill-in, and narrative reordering.

frame-level recognition or caption matching, these tasks require recovering the underlying event structure of a video.

Recent evidence suggests that current models often succeed without such structure. Performance drops sharply when evaluation moves from single-event to multi-event settings [2], and models frequently perform near chance on event reordering tasks where humans achieve near-perfect accuracy [3]. More fundamentally, models can answer a substantial portion of video QA questions without accessing visual input [15], and show limited sensitivity to temporal perturbations such as frame shuffling [10, 35]. Timestamp prediction models often output similar time ranges across unrelated videos, indicating reliance on language patterns rather than visual grounding [55]. These observations indicate that current evaluation protocols often conflate temporal understanding with language priors and world knowledge.

We argue that this gap stems from a mismatch between current post-training objectives and structured temporal understanding. Most approaches rely on text-supervised objectives, where models are trained to generate captions or answers aligned with language distributions. Such objectives allow models to exploit language shortcuts rather than grounding predictions in visual temporal evidence [15, 10]. Meanwhile, existing temporal learning signals remain fragmented across isolated skills. Prior work has explored temporal grounding [42, 53, 47, 34], event prediction [9, 40], and temporal reasoning for QA [16], while classical self-supervised tasks such as VCP [33], OPN [26], and temporal transformation recognition [24] were designed for representation learning rather than post-training MLLMs. However, no existing objective explicitly trains models to jointly recover event boundaries, intra-event progression, and inter-event relations as a unified capability.

To address this limitation, we propose *EventLens*, a vision-centric reinforcement learning framework that learns video understanding through *event-structure recovery*. The key idea is to train models not to match language descriptions, but to recover temporal structure from visual evidence. We operationalize this by deliberately perturbing temporal structure through changes in granularity, playback direction, and event order, and training the model to recover the original structure. These recovery tasks are inherently vision-centric and admit deterministic verification signals, enabling reinforcement learning without additional human preference labels or learned reward models.

To systematically instantiate this framework, we introduce a three-level temporal ontology consisting of thematic phases, semantic events, and atomic action units. Based on this ontology, we derive three complementary task families: event segmentation, progression discrimination, and multi-event relation reasoning via next-event prediction, fill-in, and clip reordering. We further design a two-stage training paradigm that separates task specialization from policy unification. We first train task-specific policies with GRPO under verifiable rewards, and then consolidate them into a unified model using multi-teacher on-policy distillation (MT-OPD), avoiding instability from directly mixing heterogeneous reward signals.

We evaluate EventLens on Qwen3-VL and Qwen2.5-VL backbones and compare it with supervised and reinforcement-learning post-training methods. Across both Qwen-VL generations, EventLens improves temporal grounding and event-sensitive reasoning while maintaining competitive general video understanding. Direct recovery-task evaluations show that EventLens integrates complementary segmentation, progression, and relation skills into a unified model. Learning curve analysis further demonstrates that EventLens achieves comparable downstream transfer with substantially reduced training time compared to direct mixed RL optimization.

In summary, our contributions are as follows:

- We formulate structured temporal understanding as *event-structure recovery* from visual evidence, unifying event boundaries, intra-event progression, and inter-event relations.
- We propose **EventLens**, a vision-centric reinforcement learning framework that learns video understanding through temporal perturbation and recovery with verifiable rewards.
- We introduce a training paradigm with task-specialized GRPO and multi-teacher on-policy distillation (MT-OPD), enabling stable integration of heterogeneous temporal skills.
- We demonstrate consistent improvements on temporal-structure benchmarks with controlled retention on general video QA, and improved training efficiency compared to direct mixed RL optimization.

2 Related work

Video MLLMs and RL post-training. Recent video MLLMs, including LLaVA-Video [54], VideoLLaMA3 [52], InternVL-2.5 [7], InternVideo2.5 [41], Qwen2.5-VL [5], and Qwen3-VL [4], show strong coarse-grained video understanding, but standard text-supervised post-training does not directly optimize structured temporal recovery. Recent RL methods introduce verifiable rewards for video QA, temporal perception, and grounding [16, 49, 42, 53, 47, 34]. VideoAuto-R1 [30] adapts reasoning to each query, while OneThinker [17] unifies image and video tasks through multi-task RL. R1-Track [39] applies MLLM RL to object tracking; ViewSAM [19] learns cross-view referring tracking under weak supervision. Closest to our use of proxy tasks, Visual Jigsaw [46], VideoSSR [21], ViSS-R1 [13], and VidBridge-R1 [6] train MLLMs with visual or temporal transformations. These proxy-based methods establish the value of temporal perturbations for video MLLMs, but they usually optimize one temporal skill, one perturbation family, or one downstream format. Our framework instead uses a shared temporal abstraction to generate complementary recovery tasks for event boundaries, single-event progression, and multi-event relations.

Temporal structure learning and evaluation. Supervised temporal action segmentation and generic event boundary detection partition videos into meaningful units, but they usually assume dense labels and one fixed granularity [14, 23, 50, 32, 37, 38, 25]. Event-centric benchmarks further test ordering, clue integration, causal logic, and temporal semantic sensitivity [48, 12, 31, 8, 2, 11]. Video-grounded entailment tree reasoning [29] connects commonsense QA to explicit verification of visual evidence. Existing training objectives cover parts of this space through text-supervised event learning, event prediction, or classical temporal pretext learning [9, 51, 40, 36, 43, 33, 26]. We adapt these intuitions to MLLM RL post-training by making segmentation prompt-conditioned across temporal levels and coupling it with single-event progression and multi-event relation reasoning under deterministic verifiers.

Distillation for multi-task post-training. Knowledge distillation and multiple-teacher distillation transfer complementary behaviors from specialized teachers to one student [22, 44]. On-policy distillation reduces train-inference mismatch by training on student-generated trajectories [1], and Video-OPD [27] applies this principle to temporal video grounding. We use this family of tech-

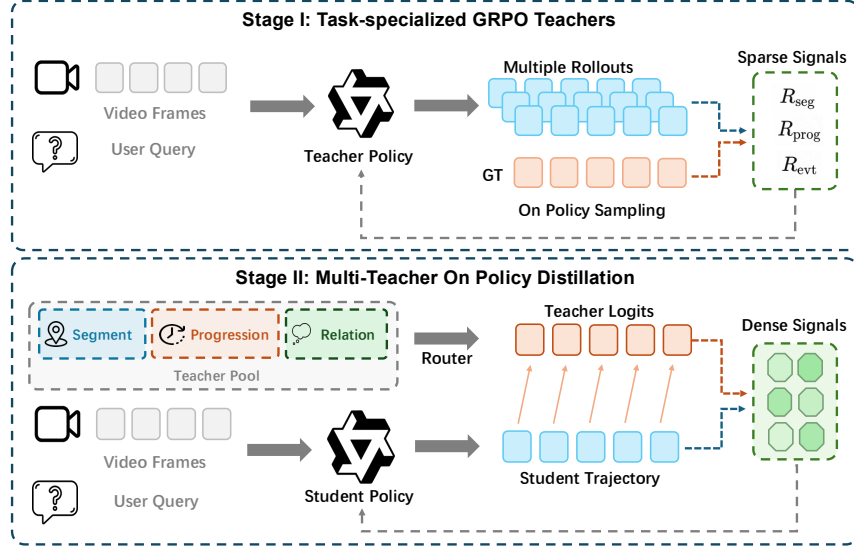


Figure 2: Overview of EventLens. EventLens adopts a two-stage training pipeline for event-structure reinforcement learning. In Stage I, deterministic rewards train task-specialized GRPO teachers for event-boundary localization, single-event progression discrimination, and multi-event relation reasoning. In Stage II, MT-OPD routes each student trajectory to the corresponding teacher and distills dense token-level teacher logits into one deployable student. The teacher is selected by recovery family, so each student trajectory is supervised by the teacher trained with the matching verifier.

niques only as the consolidation mechanism: the main distinction is that the teachers specialize in heterogeneous temporal recovery tasks whose rewards differ in scale, sparsity, and response format.

3 Method

EventLens formulates structured video understanding as recovering latent temporal structure from visual evidence. For a source video, the model must recover three complementary components of event structure: event boundaries, single-event progression, and multi-event relations. The framework constructs three event-structure recovery task families from two publicly available video corpora: **(i) level-conditioned temporal segmentation**, **(ii) single-event progression discrimination**, and **(iii) multi-event relation reasoning**. All three families are derived from a shared temporal ontology. Figure 2 summarizes how these instances feed the two-stage training pipeline. EventLens first uses these deterministic verifiers to train task-specialized GRPO teachers and then merges their skills into one deployable student with multi-teacher on-policy distillation. We denote the corresponding reward families by $\mathcal{R} = \{R_{seg}, R_{prog}, R_{evt}\}$, where R_{seg} corresponds to temporal segmentation, R_{prog} to single-event progression discrimination, and R_{evt} to multi-event relation reasoning. Each reward is computed directly from the constructed target structure, so training does not require additional human preference labels or learned reward models.

3.1 Hierarchical temporal ontology and corpus construction

Three-level temporal ontology. We organize each video with a three-level temporal ontology: L1 (thematic phase) denotes a maximal contiguous interval whose constituent events share a common high-level intent, topic, or discourse function; L2 (semantic event) denotes a coherent activity under relatively stable participants, location, and short-term goal, potentially spanning multiple shots if they preserve spatio-temporal continuity; and L3 (atomic action unit) denotes the shortest interval that still expresses a visually identifiable motion or state transition. The hierarchy moves from coarse discourse structure to fine-grained motion. All recovery tasks share this ontology, so they agree on what counts as a meaningful temporal unit.

Corpus construction. We select source videos from TimeLens-100K [53] and ET-Instruct-164K [31]. For both sources, we use only their official training partitions and exclude videos and annotations from the held-out test splits used by our evaluation benchmarks. From this parent pool, we filter videos by duration and quality. We sample frames at 1 FPS and use TransNetV2 to obtain candidate shot intervals.

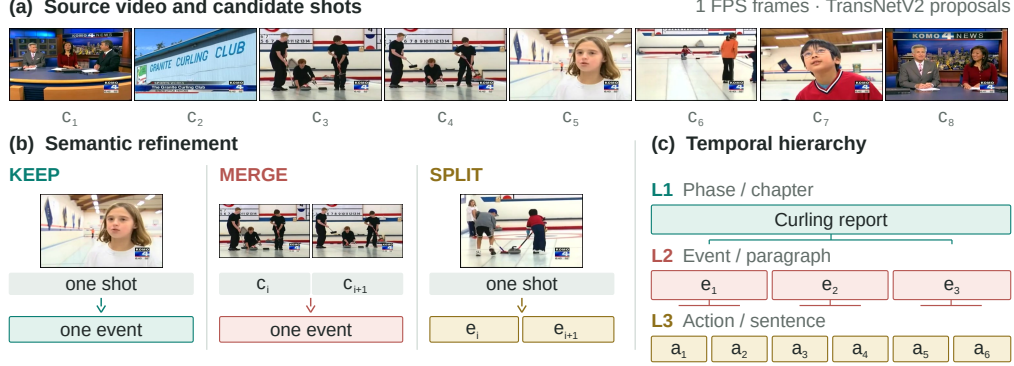


Figure 3: Scene-first construction of event structure for a curling news report. (a) Sampled frames and candidate shots. (b) KEEP retains a shot as one event, MERGE combines consecutive shots into one event, and SPLIT introduces event boundaries within a shot. (c) L1 phases contain L2 events, which contain L3 actions. Frames come from the historical report illustration; boundaries and event assignments are schematic.

Semantic refinement and hierarchy. The shot proposals are refined according to semantic continuity: consecutive shots depicting a continuous event are merged, while a shot containing multiple semantic transitions is split into distinct events (Fig. 3). The resulting semantic events define L2 intervals, each containing a sequence of nested L3 action units that capture atomic motion or state transitions. Adjacent L2 events sharing a high-level intent, topic, or discourse function are grouped into L1 phases. We enforce scene coverage, temporal validity, and parent-child containment before converting the hierarchy into recovery-task targets. Annotation details and representative examples are provided in Appendix A.1.

Recovery instance construction. From the resulting hierarchy, we obtain an ordered event chain $E = (e_1, \dots, e_M)$ at a chosen ontology level. The same filtered video can yield segmentation targets at different semantic levels, event-level clips for progression discrimination, and multi-event windows for temporal-dependency reasoning. Granularity is specified through the prompt rather than the architecture, so the same MLLM can be trained to reason over phases, events, or action units within one framework. Each instance is associated with one recovery family, which determines both the verifier and the teacher used during multi-teacher OPD.

3.2 Event-structure recovery tasks

Hierarchical temporal segmentation. This task trains the model to recover event boundaries at the semantic scale specified by the prompt. Given a video and a granularity-specifying prompt, the model outputs temporal segment boundaries at the requested level. Each training instance requests exactly one level, forcing the model to learn what constitutes a meaningful unit at that semantic scale rather than localizing a fixed text query. Because segmentation is set-valued and the number of predicted segments may differ from the target, we first compute an optimal one-to-one matching between predicted and ground-truth segments. Let $\mathcal{S}^* = \{s_j^*\}_{j=1}^{N^*}$ and $\hat{\mathcal{S}} = \{\hat{s}_i\}_{i=1}^N$ denote the ground-truth and predicted segment sets. For two temporal intervals a and b , let $\text{tIoU}(a, b) = |a \cap b| / |a \cup b|$, where $|\cdot|$ denotes duration. We obtain the matched pairs by

$$\mathcal{M}^* = \arg \max_{\mathcal{M} \in \Omega} \sum_{(i,j) \in \mathcal{M}} \text{tIoU}(\hat{s}_i, s_j^*) \quad (1)$$

where Ω is the set of all one-to-one partial matchings between $\hat{S}$ and S^* . We then score the prediction with an F1-style reward that penalizes both over- and under-segmentation:

$$R_{\text{seg}} = \frac{2P_{\text{seg}}C_{\text{seg}}}{P_{\text{seg}} + C_{\text{seg}} + \epsilon}, \quad (2)$$

$$C_{\text{seg}} = \frac{1}{N^*} \sum_{(i,j) \in \mathcal{M}^*} \text{tIoU}(\hat{s}_i, s_j^*), \quad P_{\text{seg}} = \frac{1}{N} \sum_{(i,j) \in \mathcal{M}^*} \text{tIoU}(\hat{s}_i, s_j^*).$$

where C_{seg} measures ground-truth coverage, P_{seg} measures prediction precision, and ϵ is a small constant for numerical stability. If no valid segment can be parsed from the model response, we set $R_{\text{seg}} = 0$.

Single-event progression discrimination. This task trains the model to recover the internal direction of a single event. Given an event clip, we present either the original clip or its temporally reversed counterpart and ask the model to predict the progression direction. Reversed clips are generated on the fly and require no additional annotation. Unlike classical temporal-direction pretexts on sub-second snippets [43], we sample event-level clips (typically 5–30 s), so success requires recovering coherent physical dynamics rather than relying on low-level motion artifacts. The reward is binary correctness:

$$R_{\text{prog}} = \mathbf{1}[\hat{y} = y^*] \quad (3)$$

Here $y^* \in \{\text{forward}, \text{reverse}\}$, and invalid predictions receive zero reward.

Multi-event relation reasoning. This task trains the model to recover dependencies among neighboring events. To model temporal and semantic dependencies across events, let $W = (e_1, \dots, e_m)$ denote a chronologically ordered window of events. We construct three complementary tasks from W . In next-event prediction, the model observes the visual prefix $(e_1, \dots, e_i)$ and selects which candidate event should follow. This tests whether the model can anticipate a plausible continuation from the preceding visual process. In fill-in, the model observes visual clips before and after a missing event and selects the event that best connects the two sides; this requires reasoning about both preconditions and consequences rather than only extrapolating forward. In clip reordering, the same event clips are shuffled and the model outputs their chronological permutation. For the two selection tasks, candidates contain one correct answer and hard distractors drawn from the same video or related event context, so the task cannot be solved by coarse topic matching alone. We present candidate events as visual clips with neutral identifiers rather than textual event descriptions, which keeps the verifier tied to visual temporal evidence. We treat these three tasks as one multi-event relation family with reward

$$R_{\text{evt}} = \begin{cases} \mathbf{1}[\hat{y} = y^*], & \text{for next-event prediction and fill-in,} \\ 1 - \frac{\sum_{k=1}^K |\text{pos}(k, \hat{\pi}) - \text{pos}(k, \pi^*)|}{\sum_{k=1}^K |k - (K+1-k)|}, & \text{for clip reordering,} \end{cases} \quad (4)$$

where $\hat{\pi}$ and π^* denote the predicted and ground-truth permutations, and $\text{pos}(k, \pi)$ is the position of clip k in permutation π over $K \geq 2$ clips. The first case gives exact supervision for discrete event selection. The second case uses a normalized temporal-jigsaw displacement reward, following the same motivation behind smooth order-aware supervision in prior video RL work such as VideoSSR [21], so partially correct orderings still provide dense learning signal. Malformed selections or permutations receive zero reward.

3.3 Task-specialized GRPO teachers

All event-structure recovery tasks emit deterministic rewards computed directly from task construction, making GRPO [20] a suitable optimizer for task-specialized teacher training. We train one teacher for each reward family $t \in \mathcal{T} = \{\text{seg}, \text{prog}, \text{evt}\}$. For an input x from task family t , the teacher samples a group of responses, the verifier R_t scores each response, and GRPO updates the policy with the standard group-normalized, KL-regularized objective. We summarize the teacher update with one compact objective:

$$\mathcal{J}_{\text{GRPO}}^{(t)} = \mathbb{E} \left[\frac{1}{G} \sum_{i=1}^G \rho_i^{(t)}(\theta_t) \hat{A}_i^{(t)} - \beta D_{\text{KL}}(\pi_{\theta_t} \| \pi_{\text{ref}}) \right] \quad (5)$$

where the expectation is over $x \sim \mathcal{D}_t$ and response groups $o_i \sim \pi_{\theta_{t,\text{old}}}(\cdot | x)$; $\rho_i^{(t)}(\theta_t) = \pi_{\theta_t}(o_i | x) / \pi_{\theta_{t,\text{old}}}(o_i | x)$, and $\hat{A}_i^{(t)}$ is the group-normalized advantage computed from verifier scores $R_t(o_i, x)$. For task-composition ablations, we additionally train mixed-GRPO variants by sampling multiple reward families in one run. Because segmentation, single-event progression, and multi-event relation rewards have different scales and variance profiles, these mixed-GRPO baselines use EMA-GRPO [17], which maintains task-wise exponential moving averages of reward statistics for advantage normalization.

3.4 EventLens teacher consolidation

Directly optimizing a mixed GRPO objective couples all reward families inside the same high-variance policy-gradient update. EventLens instead distills the task-specialized teachers into a single student with on-policy distillation [1], building on classical knowledge distillation [22] and multiple-teacher distillation [44]. For an input x sampled from task-family pool $\mathcal{D}_t$, the current student first generates an on-policy response. We then route the same trajectory to the corresponding teacher: segmentation examples use the segmentation teacher, single-event progression examples use the progression teacher, and next-event, fill-in, and reordering examples use the multi-event relation teacher. The routed teacher provides top- k next-token distributions on the student trajectory, and the student minimizes a sparse forward KL on this support:

$$\mathcal{L}_{\text{MT-OPD}} = \mathbb{E}_{t \sim q, x \sim \mathcal{D}_t, y \sim \pi_{\theta}(\cdot | x)} \left[\hat{D}_{\text{KL}}^{(k)}(\pi_{T_t} \| \pi_{\theta}; x, y) \right], \quad (6)$$

where q_t is the task-family sampling probability and $\hat{D}_{\text{KL}}^{(k)}$ denotes the top- k teacher-student KL along the student trajectory. Both teacher and student probabilities are renormalized over the teacher top- k set before the KL is computed. This trains the student on its own trajectories, reducing the train-inference mismatch of offline teacher-output distillation while preserving each teacher’s specialization in one deployable model.

4 Experiments

4.1 Experimental setup

Implementation details. We build on Qwen3-VL-4B and Qwen3-VL-8B [4]. All post-training uses the proxy-task data described in Sec. 3.1. Task-specialized teachers are trained with EMA-GRPO using learning rate 5×10^{-7} , global batch size 64, rollout number $G=8$, and KL coefficient 1×10^{-2} . The final EventLens student is trained with multi-teacher OPD on the same mixed proxy-task pool, with one student rollout per prompt, top- $k=10$ teacher distributions, and OPD KL coefficient 1.0. Training and evaluation use the same video budget: 256 frames and 256×256 maximum pixels. For each benchmark, inference settings are matched across the pretrained and post-trained Qwen3-VL variants, using greedy decoding. Temporal-grounding frame sampling is specified separately in Appendix F.

Benchmarks and baselines. We organize evaluation by role rather than by dataset taxonomy. Matched temporal grounding tests whether EventLens transfers to explicit boundary localization; structured reasoning tests whether the same training helps event-level inference; and general and long-video QA serve as retention checks. Dataset names, metrics, and grouped results are reported in Tables 1 and 2. For each Qwen3-VL backbone, we compare the pretrained checkpoint, two text-supervised post-training variants (SFT-Desc and SFT-DVC), and EventLens under the same inference protocol. Appendix F gives the temporal grounding prompt, timestamp parser, and metric implementation, and Appendix G provides a same-scale reference point against a recent video post-training method. We therefore interpret Tables 1 and 2 asymmetrically: temporal grounding and event-sensitive reasoning are target evaluations, whereas broad and long-video QA benchmarks are retention diagnostics rather than the primary optimization target.

4.2 Main results

EventLens delivers targeted temporal gains. Tables 1 and 2 show that EventLens is most reliable when evaluation requires explicit event localization. Across both Qwen3-VL backbones, it is the best

Table 1: Temporal grounding and complex reasoning under the shared 256-frame protocol. Temporal grounding uses mIoU; reasoning columns use official benchmark scores. Bold marks the best score within each backbone group.

Model	Temporal Grounding			Complex Reasoning			
	QV Highlights	Activity Net	Charades STA	Video Holmes	Video MMMU	Video TT	E.T. Bench
Qwen3-VL-4B Backbone							
Qwen3-VL-4B [4]	54.31	41.99	43.64	39.96	55.44	36.90	34.55
+ SFT-Desc	55.63	42.18	43.52	41.26	55.00	41.70	32.57
+ SFT-DVC	57.13	42.36	40.18	41.97	56.22	41.20	37.78
+ EventLens	58.50	45.76	47.99	40.28	55.56	38.80	41.47
Qwen3-VL-8B Backbone							
Qwen3-VL-8B [4]	50.03	39.96	44.79	39.96	60.11	40.00	32.77
+ SFT-Desc	51.44	39.70	44.80	42.24	58.00	41.80	33.67
+ SFT-DVC	57.61	43.70	42.56	42.57	57.33	42.00	39.45
+ EventLens	59.12	46.35	48.25	40.07	59.44	39.60	38.40

Table 2: General and long-video QA under the shared 256-frame protocol. Columns use official benchmark scores. Bold marks the best score within each backbone group.

Model	General Video QA				Long Video QA		
	MV Bench	Temp Compass	AoT Bench	Vino Ground	Video MME	MLVU	LongVideo Bench
Qwen3-VL-4B Backbone							
Qwen3-VL-4B [4]	62.55	74.96	53.64	37.00	65.00	56.77	60.70
+ SFT-Desc	64.62	76.35	54.39	41.40	64.30	56.77	61.30
+ SFT-DVC	64.70	75.93	54.12	40.20	65.00	58.17	62.10
+ EventLens	63.90	75.12	53.99	45.00	65.30	58.37	61.60
Qwen3-VL-8B Backbone							
Qwen3-VL-8B [4]	64.08	76.74	54.84	40.20	66.70	57.97	59.90
+ SFT-Desc	66.25	77.39	54.87	41.00	67.10	55.58	60.00
+ SFT-DVC	66.07	77.81	55.12	41.80	67.20	57.97	60.10
+ EventLens	65.42	77.17	54.55	47.20	67.50	56.77	60.60

variant in every temporal-grounding column, while structured reasoning gains concentrate on event-sensitive benchmarks such as E.T. Bench and VinoGround. On broad QA benchmarks, EventLens is not uniformly better than SFT baselines, which is expected because these benchmarks mix temporal reasoning with caption-style semantics, object recognition, and answer-format alignment. The important observation is that EventLens preserves comparable broad-QA performance while producing larger gains on the temporally targeted evaluations.

Table 3: Comparison with existing video post-training methods on Qwen2.5-VL-7B. Charades-STA and QVHighlights report mIoU; VinoGround reports the Text score and E.T. Bench the average score.

Model	Charades STA	QV Highlights	Vino Ground	E.T. Bench	MV Bench	LongVideo Bench
Qwen2.5-VL-7B [5]	38.34	27.66	29.40	30.96	63.38	59.20
Video-R1-7B [16]	28.80	22.33	31.20	29.50	60.27	56.20
EventLens-7B	42.70	43.79	41.00	35.85	64.00	58.60

This pattern extends to the Qwen2.5-VL-7B comparison with other video post-training methods (Table 3). The gains again concentrate on temporal grounding and event-sensitive reasoning, supporting event-structure recovery as a transferable training objective across the two Qwen-VL generations. Additional comparisons are reported in Appendix G.

4.3 Ablation studies

Table 4: Teacher consolidation and task-composition ablation on Qwen3-VL-4B. R1–R3 denote segmentation, progression, and relation teachers. Full MT-OPD is EventLens with all three teachers; the single-teacher oracle selects the best individual score per metric.

Variant	TimeLens avg mIoU	E.T. Bench Avg	Video MME Avg	Vino Ground Score	Video TT Score	Holmes Acc	MLVU Avg	LongVideo Bench Avg
Base	46.65	34.55	65.00	37.00	36.90	39.96	56.77	60.70
R1 teacher	50.84	40.34	65.30	39.20	38.20	40.12	58.76	61.10
R2 teacher	50.76	40.94	65.40	45.40	39.50	40.61	59.76	61.80
R3 teacher	50.96	39.87	65.00	37.80	37.20	39.19	57.57	61.60
Single-teacher oracle	50.96	40.94	65.40	45.40	39.50	40.61	59.76	61.80
MT-OPD (R1+R2)	50.73	41.44	65.40	45.00	39.50	40.72	59.36	61.70
Full MT-OPD	50.75	41.47	65.30	45.00	38.80	40.28	58.37	61.60

Teacher specialization is complementary. Table 4 separates the task-specialized teachers before they are compressed into one student. The single-teacher rows do not reveal a universal best teacher: different skills lead on target temporal grounding, retention, and reasoning metrics. The teacher oracle is therefore an upper envelope, not a deployable policy. MT-OPD recovers most of this envelope in one student, which supports the consolidation objective. Adding the relation teacher keeps temporal-grounding strength and slightly improves the event-centric reasoning average, but does not dominate every retention metric.

Table 5: Direct validation of recovery capabilities on Qwen3-VL-4B. R1–R3 denote segmentation, progression, and relation scores.

Variant	R1 Seg.	R2 Prog.	R3 Rel.
Base	0.399	0.550	0.555
R1+R2	0.589	0.680	0.574
R1+R3	0.592	0.590	0.708
R2+R3	0.444	0.700	0.705
Full MT-OPD	0.588	0.685	0.706

Direct validation separates the skills. Omitting each recovery objective reduces its corresponding capability. Adding R3 to R1+R2 improves relation reasoning while preserving segmentation and progression, showing that inter-event dependencies complement boundary and intra-event skills. The full student stays within 0.015 of the best score in each column. MT-OPD thus retains complementary recovery capabilities in one model (individual teachers in Appendix E).

MT-OPD mainly changes the optimization pathway. Direct Mixed-GRPO exposes one policy to heterogeneous rewards whose scales, sparsity patterns, and output formats differ across segmentation, progression discrimination, and event-relation reasoning. EventLens instead lets each teacher first solve a cleaner task-specific RL problem, then trains the student from dense top- k token distributions on its own trajectories. This converts sparse verifier outcomes into smoother token-level training signals, providing a mechanism for the faster student-stage transfer in Fig. 4: MT-OPD reaches a comparable transfer regime after roughly 13 hours, while direct Mixed-GRPO needs roughly 37 hours.

4.4 Analysis: Transfer and Failure Modes

Why event-structure recovery transfers. EventLens transfers across tasks because its proxy supervision does not teach a particular downstream output format; it teaches the model to bind textual decisions to visually grounded temporal units. Segmentation, progression, and relation tasks use interval lists, direction judgments, and choices or permutations, yet transfer to timestamp localization, event-level QA, and long-video QA. The common operation is boundary-sensitive temporal comparison: the model must identify when an event starts and ends, decide whether a state transition is plausible, and select the neighboring event that best preserves the visual process. This inductive bias explains why gains are most stable on temporal grounding and event-sensitive reasoning, where success depends on localized temporal evidence.

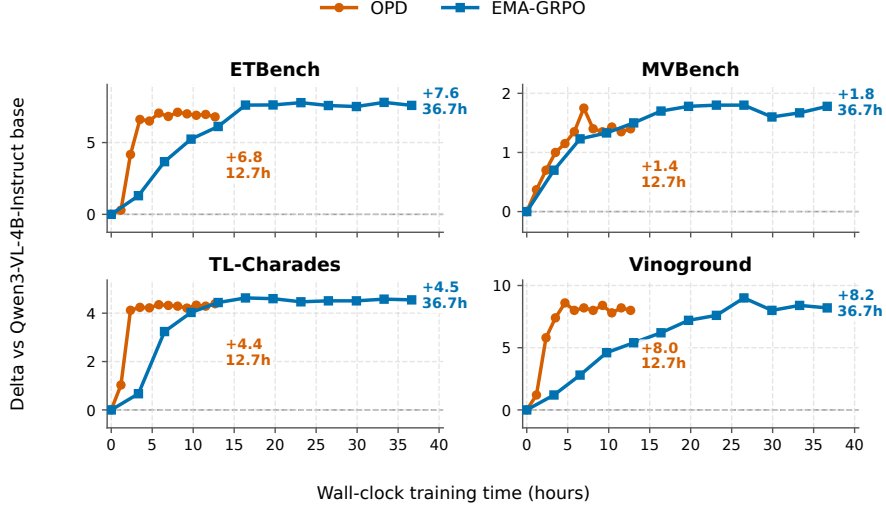


Figure 4: Wall-clock efficiency of EventLens versus direct Mixed-GRPO on four representative transfer metrics. The curves track MT-OPD student optimization and direct Mixed-GRPO training.

Table 6: Robustness to timestamp perturbation in event supervision on Qwen3-VL-4B. Training uses 100 GRPO steps and 2,000 samples per recovery family; percentages denote perturbed chains. Vinoground reports the Group score.

Training annotations	VinoGround Group	MVBench	LongVideo Bench	Charades mIoU	QVHighlights mIoU	E.T. Bench Avg.
Instruction base	37.00	62.55	60.70	43.64	54.31	34.55
Clean	39.00	63.38	61.20	45.70	58.16	38.20
10% jitter	38.60	63.20	61.30	45.61	57.99	37.92
20% jitter	38.60	63.40	61.00	45.42	57.58	37.71

Supervision robustness. EventLens remains robust to moderate perturbations of the automatically reconstructed temporal supervision. With 20% perturbation, all six metrics remain above the instruction baseline and close to clean training (Table 6), suggesting that the model learns event-level temporal structure rather than relying on exact boundary annotations.

Failure modes and trade-offs. In temporal grounding, EventLens can still confuse adjacent events when visual states change gradually, or when a query describes a high-level activity spanning several fine-grained actions. In multi-event relation tasks, visually similar repeated actions remain difficult because the correct answer may depend on subtle context rather than the appearance of an individual clip. These cases suggest that EventLens improves sensitivity to event boundaries and local temporal dependencies, but does not solve all long-range narrative reasoning. Appendix I reports the complete capability-specific results alongside qualitative cases.

5 Conclusion

We present *EventLens*, a vision-centric reinforcement learning framework for structured temporal video understanding. Instead of learning from caption-style supervision alone, EventLens trains models to recover event structure from visual evidence, including event boundaries, progression, and inter-event relations. By combining verifiable event-structure recovery tasks with task-specialized RL and policy consolidation, EventLens achieves consistent gains on temporal grounding and structured temporal reasoning while preserving broad video understanding ability. More broadly, our results suggest that recovering event structure from visual evidence provides an effective training principle for advancing temporal understanding in video MLLMs, beyond treating temporal reasoning as a byproduct of language supervision.

Acknowledgments and Disclosure of Funding

This work was supported in part by the Beijing Major Science and Technology Project under Contract no. Z251100008425023.

References

- [1] R. Agarwal, N. Vieillard, Y. Zhou, P. Stanczyk, S. Ramos Garea, M. Geist, and O. Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, volume 2024, pages 21246–21263, 2024.
- [2] D. Ahn, Y. Choi, H. Choi, S. Cho, S. Kim, and J. Choi. What happens when: Learning temporal orders of events in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2786–2796, 2026. doi: 10.1109/WACV61042.2026.00273.
- [3] P. Bai, T. Wu, J. Sun, X. Liu, H. Huang, and R. He. MVPBench: A multi-video perception evaluation benchmark for multi-modal video understanding. *arXiv:2603.22756*, 2026.
- [4] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, and K. Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.
- [5] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [6] X. Chen, Y. Zhang, Y. Guan, W. Lin, Z. Wang, B. Zeng, Y. Shi, S. Yang, Q. Liu, P. Wan, and L. Wang. VidBridge-RL: Bridging QA and captioning for RL-based video understanding models with intermediate proxy tasks. In *The Fourteenth International Conference on Learning Representations*, volume 2026, pages 116434–116461, 2026.
- [7] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, L. Gu, X. Wang, Q. Li, Y. Ren, Z. Chen, J. Luo, J. Wang, T. Jiang, B. Wang, C. He, B. Shi, X. Zhang, H. Lv, Y. Wang, W. Shao, P. Chu, Z. Tu, T. He, Z. Wu, H. Deng, J. Ge, K. Chen, K. Zhang, L. Wang, M. Dou, L. Lu, X. Zhu, T. Lu, D. Lin, Y. Qiao, J. Dai, and W. Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.
- [8] J. Cheng, Y. Ge, T. Wang, Y. Ge, J. Liao, and Y. Shan. Video-Holmes: Can MLLM think like Holmes for complex video reasoning? In *European Conference on Computer Vision (ECCV)*, 2026.
- [9] J.-H. Cheng, Y.-H. Peng, H. Zhou, V. Wang, H. Wang, H.-W. Huang, W. Chai, H.-I. Liu, K.-M. Chen, C.-Y. Yang, Y.-L. Chen, V. Vineet, Q. Cai, and J.-N. Hwang. TEMPURA: Temporal event masked prediction and understanding for reasoning in action. In *Conference on Language Modeling (COLM)*, 2026.
- [10] D. Cores, M. Dorkenwald, M. Mucientes, C. G. M. Snoek, and Y. M. Asano. Lost in time: A new temporal benchmark for video LLMs. In *The British Machine Vision Conference (BMVC)*, 2025.
- [11] Y. Du, Y. Liu, and Q. Jin. Reversed in time: A novel temporal-emphasized benchmark for cross-modal video-text retrieval. In *ACM International Conference on Multimedia (MM)*, pages 5260–5269, 2024. doi: 10.1145/3664647.3680731.
- [12] Y. Du, K. Zhou, Y. Huo, Y. Li, W. X. Zhao, H. Lu, Z. Zhao, B. Wang, W. Chen, and J.-R. Wen. Towards event-oriented long video understanding. *arXiv:2406.14129*, 2024.

- [13] B. Fang, Y. Song, H. Sun, X. Zhang, Q. Wu, W. Wu, and A. B. Chan. ViSS-R1: Self-supervised reinforcement video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 11190–11200, 2026.
- [14] Y. A. Farha and J. Gall. MS-TCN: Multi-stage temporal convolutional network for action segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3575–3584, 2019.
- [15] B. Feng, Z. Lai, S. Li, Z. Wang, S. Wang, P. Huang, and M. Cao. Breaking down video LLM benchmarks: Knowledge, spatial perception, or true temporal understanding? *arXiv:2505.14321*, 2025.
- [16] K. Feng, K. Gong, B. Li, Z. Guo, Y. Wang, T. Peng, J. Wu, X. Zhang, B. Wang, and X. Yue. Video-R1: Reinforcing video reasoning in MLLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, Main Conference, pages 99114–99137, 2025. doi: 10.52202/085713-3310.
- [17] K. Feng, M. Zhang, H. Li, K. Fan, S. Chen, Y. Jiang, D. Zheng, P. Sun, Y. Zhang, H. Sun, Y. Feng, P. Pei, X. Cai, and X. Yue. OneThinker: All-in-one reasoning model for image and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5432–5443, 2026.
- [18] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24118, 2025.
- [19] J. Ge, X. Zhang, J. Cao, B. Liu, F. Deuser, C. Liu, G. Wenkang, S. Li, J. Shao, W. Wu, C. Feng, and I. Patras. ViewSAM: Learning view-aware cross-modal semantics for weakly supervised cross-view referring multi-object tracking. *arXiv preprint arXiv:2605.02638*, 2026.
- [20] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645 (8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.
- [21] Z. He, X. Qu, Y. Li, S. Huang, D. Liu, and Y. Cheng. VideoSSR: Video self-supervised reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26530–26540, 2026.
- [22] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.
- [23] Y. Ishikawa, S. Kasai, Y. Aoki, and H. Kataoka. Alleviating over-segmentation errors by detecting action boundaries. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2322–2331, 2021.
- [24] S. Jenni, G. Meishvili, and P. Favaro. Video representation learning by recognizing temporal transformations. In *European Conference on Computer Vision (ECCV)*, pages 425–442, 2020. doi: 10.1007/978-3-030-58604-1_26.
- [25] H. Jung, D. Kim, S. Lim, J. Son, and J. Choi. Online generic event boundary detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13741–13750, 2025.
- [26] H.-Y. Lee, J.-B. Huang, M. Singh, and M.-H. Yang. Unsupervised representation learning by sorting sequences. In *IEEE International Conference on Computer Vision (ICCV)*, pages 667–676, 2017.
- [27] J. Li, H. Yin, H. Xu, B. Xu, W. Tan, Z. He, J. Ju, Z. Luo, and J. Luan. Video-OPD: Efficient post-training of multimodal large language models for temporal video grounding via on-policy distillation. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 71914–71931. PMLR, 2026.

- [28] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo, L. Wang, and Y. Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206, 2024.
- [29] H. Liu, F. Ilievski, and C. G. M. Snoek. Commonsense video question answering through video-grounded entailment tree reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3262–3271, 2025.
- [30] S. Liu, M. Zhuge, C. Zhao, J. Chen, L. Wu, Z. Liu, C. Zhu, Z. Cai, C. Zhou, H. Liu, E. Chang, S. Suri, H. Xu, Q. Qian, W. Wen, B. Varadarajan, Z. Liu, H. Xu, F. Bordes, R. Krishnamoorthi, B. Ghanem, V. Chandra, and Y. Xiong. VideoAuto-R1: Video auto reasoning via thinking once, answering twice. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 32829–32839, 2026.
- [31] Y. Liu, Z. Ma, Z. Qi, Y. Wu, Y. Shan, and C. W. Chen. E.T. Bench: Towards open-ended event-level video-language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 32076–32110, 2024. doi: 10.52202/079017-1009.
- [32] Z. Lu and E. Elhamifar. FACT: Frame-action cross-attention temporal modeling for efficient action segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18175–18185, 2024.
- [33] D. Luo, C. Liu, Y. Zhou, D. Yang, C. Ma, Q. Ye, and W. Wang. Video cloze procedure for self-supervised spatio-temporal learning. *AAAI Conference on Artificial Intelligence*, 34(07): 11701–11708, 2020. doi: 10.1609/aaai.v34i07.6840.
- [34] F. Luo, S. Lou, C. Chen, Z. Wang, C. Li, W. Shen, J. Guo, P. Li, M. Yan, J. Zhang, F. Huang, and Y. Liu. MUSEG: Reinforcing video temporal understanding via timestamp-aware multi-segment grounding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35549–35561. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1644.
- [35] H. Ok and J. Lee. TempCore: Are video QA benchmarks temporally grounded? *arXiv preprint arXiv:2509.01167*, 2025.
- [36] L. C. Pickup, Z. Pan, D. Wei, Y. Shih, C. Zhang, A. Zisserman, B. Schölkopf, and W. T. Freeman. Seeing the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2035–2042, 2014.
- [37] M. Z. Shou, S. W. Lei, W. Wang, D. Ghadiyaram, and M. Feiszli. Generic event boundary detection: A benchmark for event segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8075–8084, 2021.
- [38] J. Tang, Z. Liu, C. Qian, W. Wu, and L. Wang. Progressive attention on multi-level dense difference maps for generic event boundary detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3355–3364, 2022.
- [39] B. Wang, W. Li, and J. Ge. R1-Track: Direct application of MLLMs to visual object tracking via reinforcement learning. *arXiv preprint arXiv:2506.21980*, 2025.
- [40] H. Wang, H. Liu, X. Liu, C. Du, K. Kawaguchi, Y. Wang, and T. Pang. Fostering video reasoning via next-event prediction. In *International Conference on Learning Representations (ICLR)*, volume 2026, pages 31524–31570, 2026.
- [41] Y. Wang, X. Li, Z. Yan, Y. He, J. Yu, X. Zeng, C. Wang, C. Ma, H. Huang, J. Gao, M. Dou, K. Chen, W. Wang, Y. Qiao, Y. Wang, and L. Wang. InternVideo2.5: Empowering video MLLMs with long and rich context modeling. *arXiv:2501.12386*, 2025.
- [42] Y. Wang, Z. Wang, B. Xu, Y. Du, K. Lin, Z. Xiao, Z. Yue, J. Ju, L. Zhang, D. Yang, X. Fang, Z. He, Z. Luo, W. Wang, J. Lin, J. Luan, and Q. Jin. Time-R1: Post-training large vision language model for temporal video grounding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, Main Conference, pages 83330–83364, 2025. doi: 10.52202/085713-2793.

- [43] D. Wei, J. J. Lim, A. Zisserman, and W. T. Freeman. Learning and using the arrow of time. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8052–8060, 2018.
- [44] C. Wu, F. Wu, and Y. Huang. One teacher is enough? pre-trained language model distillation from multiple teachers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4408–4413, 2021. doi: 10.18653/v1/2021.findings-acl.387.
- [45] H. Wu, D. Li, B. Chen, and J. Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 28828–28857, 2024. doi: 10.52202/079017-0907.
- [46] P. Wu, Y. Zhang, H. Diao, B. Li, L. Lu, and Z. Liu. Visual Jigsaw post-training improves MLLMs. In *International Conference on Learning Representations (ICLR)*, volume 2026, pages 93221–93245, 2026.
- [47] T. Wu, L. Yang, G. Zhan, Y. Zhang, Y. Liao, J. Li, D. Fu, L. Zhang, and L. Wang. TempR1: Improving temporal understanding of MLLMs via temporal-aware multi-task reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2756–2767, 2026.
- [48] J. Xiao, X. Shang, A. Yao, and T.-S. Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, 2021.
- [49] Z. S. Xue, R. Luo, and K. Grauman. Seeing the arrow of time in large multimodal models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, Main Conference, pages 90925–90955, 2025. doi: 10.52202/085713-3041.
- [50] F. Yi, H. Wen, and T. Jiang. ASFormer: Transformer for action segmentation. In *The British Machine Vision Conference (BMVC)*, 2021.
- [51] X. Zeng, K. Li, C. Wang, X. Li, T. Jiang, Z. Yan, S. Li, Y. Shi, Z. Yue, Y. Wang, Y. Wang, Y. Qiao, and L. Wang. TimeSuite: Improving MLLMs for long video understanding via grounded tuning. In *International Conference on Learning Representations (ICLR)*, volume 2025, pages 38057–38081, 2025.
- [52] B. Zhang, K. Li, Z. Cheng, Z. Hu, Y. Yuan, G. Chen, S. Leng, Y. Jiang, H. Zhang, X. Li, P. Jin, W. Zhang, F. Wang, L. Bing, and D. Zhao. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv:2501.13106*, 2025.
- [53] J. Zhang, T. Wang, Y. Ge, Y. Ge, X. Li, and L. Wang. TimeLens: Rethinking video temporal grounding with multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10419–10429, 2026.
- [54] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.
- [55] H. Zhao, G.-P. Ji, R. Yan, H. Xiong, and Z. Li. VideoExpert: Augmented LLM for temporal-sensitive video understanding. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(5):5893–5908, 2026. doi: 10.1109/TCSVT.2026.3653742.
- [56] J. Zhou, Y. Shu, B. Zhao, B. Wu, Z. Liang, S. Xiao, M. Qin, X. Yang, Y. Xiong, B. Zhang, T. Huang, and Z. Liu. MLVU: Benchmarking multi-task long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13691–13701, 2025.

Appendix

This appendix provides additional implementation details, evaluation protocols, and qualitative analyses that complement the main paper. We first describe the construction of the event-structure proxy-task pool, including data sources, recovery views, reward parsing rules, and the multi-teacher OPD implementation (§A–C). We then report training hyperparameters and the detailed temporal grounding evaluation protocol for reproducibility (§D–F). Next, we provide additional empirical analyses, including comparisons with related video post-training methods, qualitative case studies, and full efficiency curves beyond the representative metrics shown in the main text (§G–J). Finally, we discuss the current limitations of EventLens and directions for future work (§K).

A Proxy-task data details

The post-training pool converts source videos into three event-structure recovery views that mirror the main EventLens tasks. Table 7 summarizes what each view asks the model to recover and which deterministic target is used for verification. All instances are built from the official training partitions of TimeLens-100K [53] and ET-Instruct-164K [31]; videos and annotations from held-out test splits of TimeLens, E.T. Bench, or any other evaluation benchmark are excluded from the post-training pool. Thus, naming overlap between a training source and an evaluation benchmark does not imply train–test sharing. For reproducibility, each stored example contains a prompt, one or more video frame lists, a structured target, metadata, and a recovery-family label used only to select the matching verifier and teacher.

Table 7: Event-structure recovery views used for post-training.

Recovery view	Constructed recovery task	Deterministic target
Event-boundary recovery	Recover phase-, event-, or action-level time spans requested by the prompt.	Event intervals; balanced by level and domain
Single-event progression	Decide whether an event-level clip is played forward or backward.	Direction label; balanced by playback direction and domain
Multi-event relation	Recover the next event, a missing middle event, or the order of shuffled clips.	Choice or ordering; balanced by relation subtask and domain

The proxy-task pool is intentionally built from a diverse set of video domains rather than from a single narrow source. Fig. 5 visualizes the domain composition of the input videos before conversion into segmentation, single-event progression, and multi-event relation instances. This diversity is important because the proxy tasks manipulate event structure rather than dataset-specific labels: the same verifier design can supervise event boundaries, single-event progression, and multi-event relations across instructional, human-activity, object-centric, and scene-level videos.

For video inputs, we use offline frame lists rather than decoding videos during training. The experiment builder derives frame lists from trusted 2 FPS frame caches and applies a duration-dependent frame policy: clips shorter than 60 seconds are kept at 2 FPS, while longer clips are sampled at 1 FPS, followed by a uniform cap of 256 frames. The same 256-frame and 256×256 pixel budget is used for training and evaluation.



Figure 5: Domain composition of the input videos used to construct proxy-task instances.

A.1 Hierarchy construction and supervision assessment

Figure 3 uses a curling news report to illustrate how candidate shots are organized into a temporal hierarchy. The sampled frames span a studio presentation, the curling-club exterior, on-ice demonstrations, and player interviews. Camera changes supply candidate boundaries; semantic refinement determines which consecutive or within-shot intervals form coherent L2 events.

Table 8: Semantic refinement and temporal hierarchy in the curling-report schematic. Frame identifiers refer to Figure 3; event boundaries and assignments are illustrative.

Decision	Report content	Semantic interpretation
KEEP	Interview view c_5 : one speaker continues within a single shot.	Retain the candidate interval as one L2 event when its semantic focus remains unchanged.
MERGE	Adjacent views c_3 and c_4 show the same group performing a curling demonstration.	Combine consecutive candidates c_i and c_{i+1} into one L2 event when a camera cut preserves the activity.
SPLIT	A single rink shot can cover successive segments, such as preparing a delivery and demonstrating the play.	Place an internal boundary at a change of semantic focus, producing separate L2 events e_i and e_{i+1} .
L1 phase	The complete curling report, from the studio presentation to the return to the anchors.	Group related L2 events under the shared topic, "Curling report," while preserving their temporal order.
L2 event	Club context, on-ice demonstrations, and player interviews provide distinct event-level content.	Describe each coherent report segment with an event interval; e_1 , e_2 , and e_3 illustrate the event sequence.
L3 action	Local actions, such as releasing a stone or sweeping, occur inside an on-ice demonstration.	Keep each action interval inside its parent L2 event, and each L2 event inside the enclosing L1 phase.

Temporal organization. The hierarchy separates a report’s overall topic, its coherent semantic segments, and the local actions within each segment. A camera cut can occur inside one event, while a single shot can span an event transition. Table 8 therefore ties refinement to continuity or a change of semantic focus. Once event intervals are established, L3 actions remain temporally contained in their parent events, and related L2 events are aggregated into an L1 phase.

Construction implementation. We sample annotation frames at 1 FPS and use TransNetV2 to obtain candidate shot intervals. Semantic refinement retains a shot as one event when its focus remains unchanged, merges consecutive shots that preserve event continuity, and splits a shot at semantic transitions. The resulting L2 events contain L3 action units that capture atomic motion or state transitions; adjacent L2 events sharing a high-level intent, topic, or discourse function are grouped into L1 phases. Before task conversion, we check scene coverage, temporal validity, parent–child containment, temporal ordering, and non-overlap. Extremely short, ambiguous, or overly fragmented intervals are discarded. Each segmentation prompt requests one temporal level, and progression clips preserve coherent event-level state transitions. For relation recovery, distractors are sampled preferentially from the same video or semantically related neighboring contexts; candidates are visual clips with neutral identifiers rather than textual descriptions.

Model-assisted audit. We inspect 100 sampled event chains using three MLLM judges and four criteria: caption grounding, boundary acceptability, hierarchy consistency, and proxy answerability. Table 9 reports per-judge and majority-voted acceptance; 87% of chains satisfy all four criteria under majority voting.

Table 9: Model-assisted quality audit of 100 event chains. Columns report acceptance for each criterion and for all four criteria jointly.

Judge	Caption grounded	Boundary acceptable	Hierarchy consistent	Proxy answerable	All four passed
GPT-5.5	97%	91%	97%	87%	83%
Claude-4.8	94%	89%	95%	88%	73%
Gemini-2.5-Pro	99%	97%	100%	82%	82%
Three-judge majority	98%	93%	98%	91%	87%

Timestamp perturbation. Table 6 reports the Qwen3-VL-4B experiment using 2,000 samples per recovery family and 100 GRPO steps. We select 10% or 20% of chains and perturb their temporal coordinates while preserving interval duration, event text, and sample identity. Percentages denote the selected-chain fraction, not boundary displacement magnitude. This protocol assesses timestamp imprecision in recovery supervision.

B Reward and parsing details

The reward parser follows the same three event-structure recovery families used in Section 3.2: event-boundary recovery, single-event progression, and multi-event relations. This section specifies how model text is converted into the verifier inputs in Eqs. 2 and 4. Format validity is logged for analysis, but the training reward is always the task verifier defined by the constructed target structure.

Event-boundary recovery. Segmentation responses are parsed from the `<events>` block as a list of $[s, e]$ intervals. Invalid intervals, missing tags, malformed bracket formats, and repeated event tags receive zero reward because no reliable boundary set can be recovered. Valid predictions are de-duplicated with one-dimensional NMS and matched to ground-truth intervals with Hungarian matching. The matched intervals are then scored by the F1-IoU reward in Eq. 2, so the same parser supports phase-, event-, and action-level segmentation prompts.

Single-event progression. Progression-discrimination responses are parsed as a multiple-choice decision over the forward and reverse playback options. The parser first extracts the final letter inside `<answer>` and falls back to the last standalone option letter if the tag is missing. The reward is one only when the parsed option matches the constructed playback direction and zero otherwise, matching R_{prog} in Section 3.2.

Multi-event relations. Next-event prediction and fill-in use the same multiple-choice parser as progression discrimination, but their target option is derived from the hidden event in the local event window. Clip reordering instead parses a numeric permutation from `<answer>`. Responses with no parseable in-range clip index receive zero reward. For parseable but incomplete permutations, duplicate indices are removed, overlong predictions are truncated to the target length, and missing target indices are appended before scoring so that every valid partial order maps to a complete permutation. We then compute the normalized temporal-jigsaw displacement in Eq. 4, which gives full credit to the exact order and partial credit when clips remain near their correct temporal positions.

C Multi-teacher OPD implementation

The implementation of multi-teacher OPD follows a simple routed distillation loop. Unlike GRPO teacher training, OPD uses one student rollout per prompt and does not compute reward groups, advantages, online filtering, or critic values. After the student generates an on-policy response, the corresponding teacher scores the same student trajectory with top- k next-token log probabilities. The student then gathers its log probabilities at the teacher top- k token ids and minimizes a sparse forward KL after renormalizing both distributions on this support.

1. Sample prompts from the mixed proxy-task pool.
2. Generate responses $y \sim \pi_{\theta}(\cdot \mid x)$ with the current student.
3. Route each example to the teacher trained for its recovery family.

4. Query the routed teacher for top- k token distributions along y .
5. Update the student with the sparse OPD KL loss.

Teacher routing follows the same task grouping as the data pipeline: segmentation examples are routed to the segmentation teacher, single-event progression examples to the progression teacher, and multi-event relation examples to the relation teacher. For memory efficiency, teacher models are treated as reference modules and can be loaded, queried on their sub-batch, and offloaded sequentially. We also use task-homogeneous batching so that each mini-batch contains examples from a single coarse task group whenever possible.

D Training hyperparameters

Appendix D reports the shared implementation settings used across all Qwen3-VL backbones unless otherwise specified. The teacher-related rows correspond to task-specialized EMA-GRPO training, while the OPD-related rows correspond to the final multi-teacher student consolidation stage. We keep the video budget and decoding protocol fixed across pretrained, SFT, teacher, and EventLens variants, so that reported differences are attributable to post-training rather than inference-time frame allocation, resolution, or sampling choices.

Table 10: Training and inference settings.

Item	Value	Item	Value
Backbones	Qwen3-VL-4B / 8B	Teacher training	EMA-GRPO
Student training	Multi-teacher OPD	Teacher LR	5×10^{-7}
Teacher batch	64	GRPO rollouts	$G = 8$
Teacher KL	1×10^{-2}	OPD rollouts	$n = 1$
OPD teacher dist.	top- $k = 10$	OPD KL	1.0
Teacher routing	By recovery family	Eval. decoding	Greedy
Max frames	256	Max pixels	256×256

The one-rollout OPD setting is used because MT-OPD obtains token-level teacher distributions on the student trajectory, rather than estimating group-normalized rewards as in GRPO. The top- k teacher distribution keeps the routed multi-teacher distillation target sparse and memory-efficient while preserving high-probability alternatives for stable KL matching.

E Individual-teacher recovery capabilities

Table 11: Individual-teacher recovery scores on Qwen3-VL-4B.

Variant	R1	R2	R3
	Segmentation	Progression	Relation
Base	0.399	0.550	0.555
R1	0.587	0.570	0.580
R2	0.435	0.685	0.570
R3	0.450	0.565	0.705
Full MT-OPD	0.588	0.685	0.706

Specialization and consolidation. Each teacher produces its largest gain on its corresponding recovery task. Relative to Base, R1 improves segmentation by 0.188, R2 improves progression by 0.135, and R3 improves relation reasoning by 0.150; their gains on the other two tasks are smaller. Full MT-OPD matches or exceeds the best individual teacher in every column. Together with the task-combination results in Table 5, these results support training complementary teacher policies and consolidating their recovery capabilities into a single student.

F Temporal grounding evaluation details

We evaluate temporal grounding with the same VLMEval-style implementation for all pretrained, SFT, and EventLens variants. The three reported TimeLens subsets, QVHighlights, ActivityNet, and

Charades-STA, are loaded through the TimeLensBench interface, which converts each annotation into a TSV row containing the video id, natural-language query, ground-truth span, and video duration. Videos are supplied as structured video inputs when the evaluated backend supports video messages; otherwise the framework extracts frames with the same adaptive policy used in the rest of the paper: videos up to 60 seconds are sampled at 2 FPS, videos between 60 and 256 seconds are sampled at 1 FPS, and longer videos are uniformly capped at 256 frames. All models are decoded greedily, and no benchmark-specific demonstrations or external judges are used for these temporal grounding scores.

For the main temporal grounding results, we use the following direct prompt:

Please find the visual event described by a sentence in the video, determining its starting and ending times. The format should be: “The event happens in the start time - end time seconds”. For example: The event “person turn a light on” happens in the 24.3 - 30.4 seconds. Now I will give you the textual sentence: {query}. Please return its start time and end time in seconds.

The evaluation parser extracts the first valid pair of decimal timestamps from model outputs in either “X–Y” or “X to Y” form and discards pairs whose end time is earlier than the start time. For a predicted span $[\hat{s}, \hat{e}]$ and a reference span $[s, e]$, temporal IoU is computed as

$$\text{tIoU} = \frac{\max(0, \min(\hat{e}, e) - \max(\hat{s}, s))}{\max(\hat{e}, e) - \min(\hat{s}, s)}. \quad (7)$$

If no valid timestamp pair is found, the example receives zero IoU; malformed reference annotations are excluded from aggregation. We report mean IoU (mIoU) in the main tables and additionally log recall at IoU thresholds 0.3, 0.5, and 0.7 for reproducibility.

G Comparison with related video post-training work

Table 12: Comparisons with related video post-training methods. Charades-STA, QVHighlights, and ActivityNet report mIoU; VInoGround reports the Text score and E.T. Bench the average score. Bold marks the best score per column.

Model	Charades STA	QV Highlights	Activity Net	Vino Ground Text	E.T. Bench Avg.	MV Bench	LongVideo Bench
Qwen3-VL-8B [4]	44.79	50.03	39.96	40.20	32.77	64.08	59.90
VideoSSR-8B [21]	41.52	53.67	39.81	49.20	41.94	66.72	60.90
VideoAuto-R1-8B [30]	41.95	56.94	42.54	43.60	39.36	67.07	62.70
OneThinker-8B [17]	40.09	54.07	38.90	36.80	31.14	59.77	57.80
EventLens-4B	47.99	58.50	45.76	45.00	41.47	63.90	61.60
EventLens-8B	48.25	59.12	46.35	47.20	38.40	65.42	60.60

Table 3 reports the Qwen2.5-VL-7B comparison in the main text. Table 12 brings together the additional video post-training methods, with each model listed once across the reported benchmarks. EventLens-8B achieves the highest reported scores on Charades-STA, QVHighlights, and ActivityNet, connecting event-structure recovery to temporal grounding. Against VideoSSR-8B, EventLens improves the mean TimeLens score from 45.00 to 51.24.

H Qualitative case study

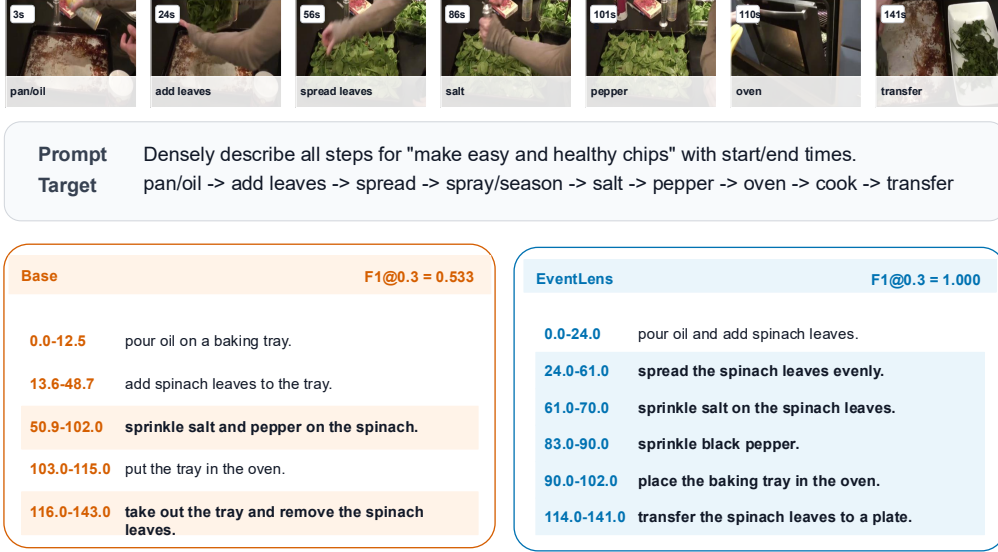


Figure 6: Qualitative E.T. Bench case on the instruction “make easy and healthy chips.” The base model recovers the coarse cooking procedure but merges several adjacent fine-grained events, while EventLens trained with MT-OPD separates the event boundaries more faithfully and obtains perfect F1@0.3 on this example.

Fig. 6 illustrates the difference between coarse procedure recognition and event-structure recovery. The base model identifies the overall recipe flow, but its segments remain too coarse: it merges the salt and pepper operations into one long interval and shifts the final transfer step toward the oven-removal phase. These errors are understandable from a captioning perspective because the predicted steps are semantically plausible, but they are harmful for dense event evaluation where adjacent visual actions must be separated in time.

EventLens with MT-OPD produces a more temporally factorized description. It groups the initial oiling and leaf placement when the visual evidence is continuous, then separates spreading, salting, peppering, oven placement, and transfer into distinct intervals. This case supports the main quantitative trend: the benefit of EventLens is not merely better recipe semantics, but improved alignment between textual event descriptions and their visual temporal boundaries. It also clarifies why some caption-style SFT baselines can remain competitive on broad reasoning scores while still missing the structured boundary precision targeted by our proxy tasks.

Fig. 7 shows that the qualitative behavior of EventLens extends beyond a single dense procedure example. In repeated action localization, the base model collapses several visually similar “folding” occurrences into one nearly full-video prediction, whereas MT-OPD recovers the repeated event granularity. In temporal event matching, the base model drifts to a tail artifact, while MT-OPD selects the later visually matched event segment. In grounded video QA, the base model chooses a plausible but incorrect tool answer, whereas MT-OPD selects the small knife and aligns the answer with the supporting frames. These examples support the role of event-structure supervision: it improves not only whether the model names the right event, but also whether the answer is grounded in the correct temporal evidence.

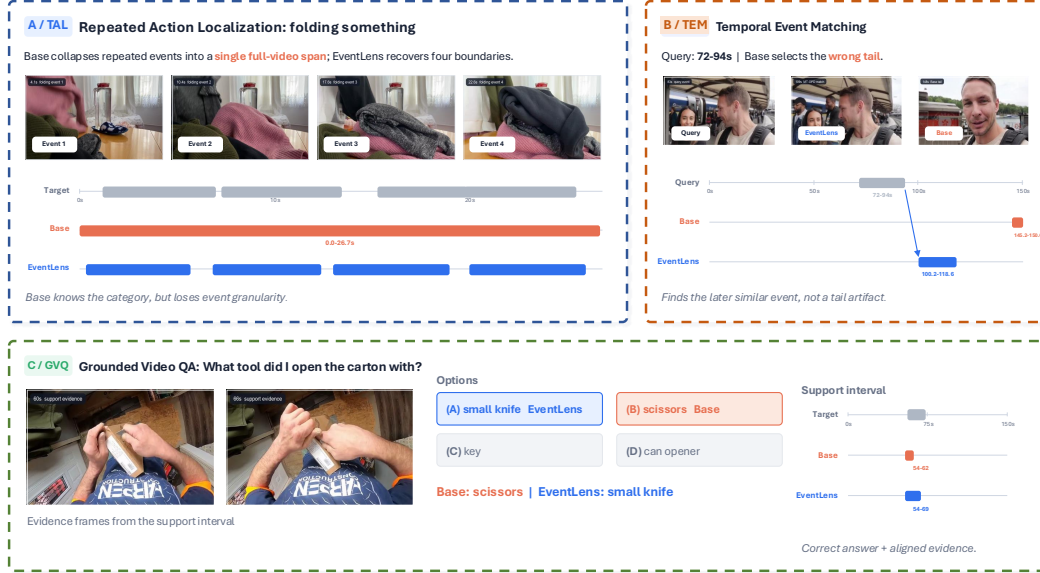


Figure 7: Additional E.T. Bench qualitative cases covering repeated action localization, temporal event matching, and grounded video QA. Across the three settings, the base model often recognizes the coarse visual category or plausible answer but misaligns it with the relevant temporal evidence, while EventLens with MT-OPD preserves finer event boundaries and support intervals.

I Capability-specific transfer and trade-offs

Table 13: Capability-level changes of EventLens relative to the corresponding Qwen3-VL instruction base. Δ denotes score-point differences for 4B and 8B models. The LongVideoBench entry retains the reported ‘Overall / TOS’ label and is separate from the aggregate score in Table 2.

Structural capability	Benchmark subset	Δ	Δ
		4B	8B
Event localization	E.T. Bench TAL/F1	+7.61	+7.87
Event grounding	E.T. Bench GND/F1	+5.14	+8.33
Temporal order	MLVU Order	+8.57	+1.43
Motion direction	TempCompass Direction	+1.06	+1.36
Text-event matching	VinoGround Text Score	+8.00	+7.00
Group-level matching	VinoGround Group Score	+2.40	+1.80
Long-range temporal ordering	LongVideoBench Overall / TOS	-4.10	-1.40
Montage plot attributes	Video-TT Plot Attribute (Montage)	-4.08	-4.08
Motion speed	TempCompass Speed	-2.84	-2.21
Anomaly recognition	MLVU Anomaly Recognition	-2.56	-2.56

Transfer aligns with the recovery objectives. Table 13 shows gains on event-sensitive capabilities at both model scales. Event localization improves by 7.61/7.87 points at 4B/8B, event grounding by 5.14/8.33 points, and text-event matching by 8.00/7.00 points. Temporal ordering, motion direction, and group-level matching also improve at both scales. These gains align with the boundary localization, intra-event progression, and inter-event relations emphasized by the three recovery objectives.

Subsets probe different information. Broad video-QA benchmarks combine temporal reasoning with static recognition, scene semantics, external knowledge, and answer-format compliance. The negative changes on montage plot attributes, motion-speed estimation, and anomaly recognition occur at both scales; these subsets additionally probe global plot organization, motion rates, and atypical event semantics. Narrative-scale ordering also calls for context beyond the finite event windows used by the current relation objective. This capability profile motivates complementing event-structure recovery with narrative-scale and fine-grained perceptual objectives.

J Additional EventLens efficiency curves

Section 4 reports four representative transfer metrics that are closely aligned with our temporal proxy tasks, so that the main text focuses on the evaluation targets most directly connected to the training objective. To make this selection transparent rather than selective, Fig. 8 reports the remaining four benchmarks from the same eight-benchmark suite. Together with Fig. 4, these curves cover the full suite without duplicating panels between the main text and appendix. The aggregate conclusion is consistent with the main text: averaged over all eight metrics, EventLens obtains a +3.39 gain after 12.7 hours, compared with +2.87 for Mixed-GRPO at its nearest wall-clock checkpoint (13.0 hours) and +3.38 for Mixed-GRPO after 36.7 hours.

The broad video QA benchmarks should be interpreted more cautiously than the targeted temporal metrics. Video-MME and Video-Holmes show small, non-monotonic changes whose magnitudes are much smaller than the gains on ETBench, TimeLens, and Vinoground. We therefore use the full-suite curves across the main text and appendix as a transparency check rather than as evidence that every benchmark improves monotonically. The main empirical claim is wall-clock efficiency: under the same mixed proxy-task pool, EventLens reaches a comparable average transfer level in roughly one third of the training time required by direct Mixed-GRPO.

K Limitations

The coverage, granularity, and semantic accuracy of automatically constructed event chains determine the quality of the recovery supervision. Our current construction audit uses model-assisted judgments, and the timestamp-perturbation experiment is a 100-step pilot. Human validation and experiments that vary event semantics or hierarchy would extend the assessment of supervision quality.

The present evaluation covers Qwen3-VL and Qwen2.5-VL under the reported video-input settings. Recovery tasks use event clips and local multi-event windows; extending them to longer narratives, denser temporal sampling, and higher-resolution motion is a natural next step.

The efficiency curves measure the MT-OPD student stage. A complete compute account also includes task-specialized teacher training and teacher-distribution queries. The consolidated student runs independently of its teachers at inference.

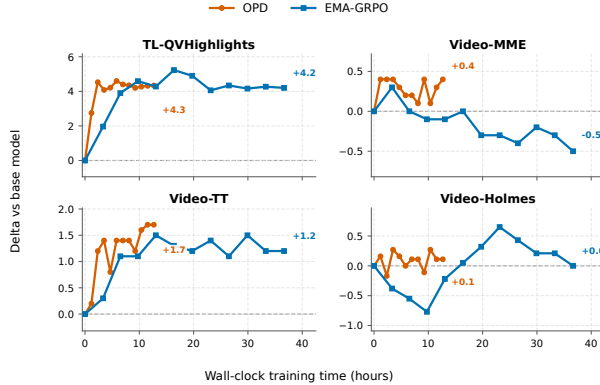


Figure 8: Additional wall-clock transfer curves complementing the four representative metrics in Fig. 4.