LEARNING PREDICTIVELY GROUNDED EVENT CHAINS FOR VIDEO EVENT PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video event prediction requires inferring an unseen continuation from visual history. An explicit event chain may be visually supported but omit predictive evidence, or appear predictive by inventing events. We define predictive grounding as the ideal conjunction of interval-level visual consistency (VC) and candidate-conditioned predictive sufficiency (PS). We introduce *Predictively Grounded Event-chain Reinforcement Learning* (PGE-RL), which optimizes verifier-relative proxies for both criteria without event-chain annotations or a task-specific supervised cold start. A shared policy first samples an option-blind timestamped history, then predicts from the video, candidates, and an immutable copy of that history. Pass 2 may revisit the video but cannot revise the chain. Frozen visual and blindfolded text verifiers supply training rewards; PS measures discriminative usefulness, not factual correctness. Five-state perception-aware credit routing assigns separate advantages to chain and prediction tokens. All verifiers are removed at inference. Reported point estimates on FutureBench, VEPBench, and AVEP favor the combined method under separate benchmark-specific training. Single-verifier ablations show an alignment–sufficiency trade-off, and chain interventions support behavioral dependence without establishing a causal mechanism. Aggregate metric inconsistencies and diagnostic aggregation remain to be reconciled with evaluation records, limiting the quantitative conclusions.

1 INTRODUCTION

Video event prediction (VEP) asks what happens after an observed clip. A multiple-choice model receives a video V and candidate futures $\mathcal{O}$, then selects a continuation from preparatory motion, object states, and ongoing interactions (Lei et al., 2020; Wang et al., 2025a). The challenge is to retain the visual details that distinguish the candidates rather than rely on a generic event script. Answer accuracy alone does not establish this: vision-language models can exploit question–answer priors or follow stale textual reasoning after visual evidence changes (Agrawal et al., 2018; Gou et al., 2026; Shi et al., 2026).

Timestamped event chains offer an inspectable interface between perception and prediction (Su et al., 2026), but can fail in two distinct ways. A visually consistent chain can be too generic to discriminate among futures; a discriminative chain can invent an event absent from the claimed interval. We call a chain *predictively grounded* when its events are supported by their video intervals (visual consistency, VC) and its explicit content provides evidence distinguishing the recorded continuation from the candidate alternatives (predictive sufficiency, PS). This is an ideal joint criterion (Figure 1); operational VC/PS scores are verifier-relative proxies, and PS is not factual verification.

We introduce *Predictively Grounded Event-chain Reinforcement Learning* (PGE-RL), training a video-language policy with GRPO without chain labels or a task-specific SFT cold start (Shao et al., 2024; DeepSeek-AI et al., 2025). Pass 1 samples a timestamped history without candidates. Pass 2 predicts from the video, candidates, and immutable history; it may revisit the video but cannot revise the chain. Frozen visual and blindfolded text verifiers assess interval support and candidate discrimination (Wang et al., 2026; Liu et al., 2024). Perception-aware credit routing (PACR) uses outcome, visual reliability, and sufficiency to assign separate chain/prediction advantages in five states. Verifiers are removed at inference.

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.

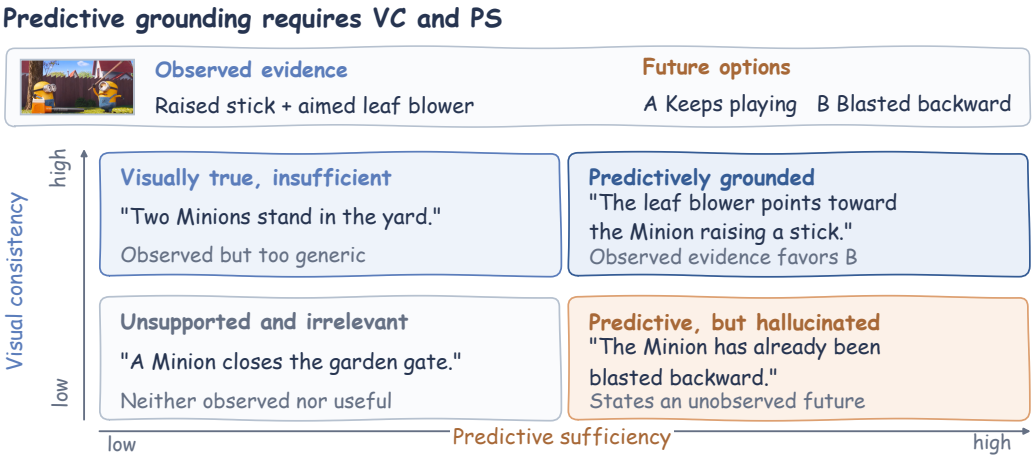


Figure 1: The ideal predictive-grounding criterion requires both visual consistency and predictive sufficiency. The examples illustrate semantic cases; automatic verifier scores only approximate these properties.

The contribution is the combination of this evidence interface and routing rule for unseen-future prediction. Event-chain scaffolds, blindfolded sufficiency, and stage-specific credit each have precedents; our comparison focuses on their concrete use together rather than priority over those pieces. The evaluation addresses three questions:

1. Do the VC and PS objectives favor different chain properties? The single-verifier ablations show a trade-off between the reported alignment and sufficiency proxies (Table 4).
2. Does prediction use the chain when the video remains available? Removing or corrupting the chain reduces reported accuracy, supporting behavioral dependence, subject to distribution-shift confounds (Table 5).
3. Does the combined method improve prediction? The recorded point estimates favor PGE-RL over the stated answer-only controls and show selected gains on AVEP under separate benchmark-specific training (Tables 1–3). Aggregate-score reconciliation and uncertainty remain unresolved, as detailed in the evaluation scope and Appendix A.6.

2 Related Work

Video event prediction. VisualCOMET predicts events and intents, VLEP formulates continuation as language-based selection, and NExT-QA evaluates temporal and causal relations among observed actions (Park et al., 2020; Lei et al., 2020; Xiao et al., 2021). FutureBench, VidEvent, and AVEP extend the evaluation or output space with reasoning depth, MLLM future-event prediction, and structured verb–participant prediction (Wang et al., 2025a; Liang et al., 2025; Su et al., 2025). Video-CoE already constructs a timestamped chain before prediction and scores timestamp-conditioned clips (Su et al., 2026). Our distinction is the candidate-exclusion and immutable-copy interface combined with per-event VC, candidate-conditioned PS gain, and five-state routing. FUTURE-L1 instead keeps latent visual state and starts from an SFT cold start (Jiang et al., 2026a).

Vision-language reasoning and grounded RL. Video and multimodal reasoning systems improve visual inspection, retrieval, or answer accuracy, while GRPO-style methods provide outcome-verifiable policy updates (Bai et al., 2025b;a; Shao et al., 2024; DeepSeek-AI et al., 2025; Feng et al., 2025; Zhang et al., 2025a; Liu et al., 2025). Grounded multimodal RL adds perception evidence to the reward: ViGoRL, PEARL, and VideoP2R separate or gate perception and downstream reasoning signals, and MoCA uses a blindfolded text reasoner to test whether perception text is sufficient for the task (Sarch et al., 2025; Zhang et al., 2026; Jiang et al., 2026b; Wang et al., 2026). MoCA's test addresses usefulness of supplied text, not whether each event is true in its claimed interval.

Conversely, visual alignment or attention rewards do not establish candidate discrimination. VisualSwap and reflective perception expose stale or unsupported reasoning (Shi et al., 2026; Wei et al., 2025). GraphThinker uses event-based video scene graphs and a visual-attention reward (Cheng et al., 2026). VideoP2R already assigns stage-specific advantages, and MoCA protects or penalizes perception tokens on failed trajectories. PACR combines outcome, interval reliability, and candidate sufficiency in five states; separate credit alone is not the novelty.

These precedents motivate our combination for unseen-future selection. Its contribution is the joint evidence interface and routing rule; verifier-relative VC/PS scores do not certify factual grounding.

3 METHOD

3.1 PREDICTIVELY GROUNDED EVENT CHAINS

A training example is $\mathcal{X} = (V, Q, \mathcal{O}, y^*)$, where V is the observed video, $\mathcal{O} = \{o_k\}_{k=1}^{K}$ contains candidate futures, and y^* indexes the recorded continuation. We treat y^* as the semantic identity of the recorded continuation rather than as its displayed option position; option order is randomized independently in the policy and verifier inputs. The task context Q appears in the documented Pass-2 and PS-verifier prompts (including benchmark question anchors); it is suppressed below where shared. Pass 1 receives neither Q nor candidates. Outcome supervision provides only $R_{\text{ans}} = \mathbb{I}[\hat{y} = y^*]$, which does not reveal whether the answer follows from the video.

We instead expose a history

$$\mathcal{E} = [E_1, \ldots, E_n], \qquad E_j = ([t_j^{\text{s}}, t_j^{\text{e}}], D_j), \tag{1}$$

where t_j^{s} and t_j^{e} are start and end times in the observed-video time coordinate, and D_j is the text description of an event claimed to occur in that interval. We require parsed events to be chronological, to have $t_j^{\text{s}} < t_j^{\text{e}}$, and to lie inside the observed duration. We define

$$\text{PG}(\mathcal{E}) \equiv \text{VC}(\mathcal{E}, V) \wedge \text{PS}(\mathcal{E}, \mathcal{O}, y^*). \tag{2}$$

Visual consistency (VC) requires every event to agree with its claimed interval and not to describe an unobserved future event; predictive sufficiency (PS) requires the explicit chain to contain evidence that distinguishes the gold future from the alternatives. This conjunction defines an ideal semantic criterion rather than an observable training label. The soft scores below are verifier-relative proxies for its two clauses. Human agreement is assessed separately (Section 4.3); neither verifier is a ground-truth annotation, and PS is not factual verification. The definition permits different decompositions of the same history rather than requiring agreement with one teacher-generated chain.

3.2 TWO-PASS GENERATION

The same policy produces each rollout in two passes with shared parameters. Pass 1 receives only the video and an event-perception prompt:

$$\mathcal{E}^{(i)} \sim \pi_\theta(\mathcal{E} \mid V, p_{\text{ec}}). \tag{3}$$

It is prompted to generate two to five timestamped events. Candidates, future frames, and labels are unavailable at sampling time. The parser checks matched event/termination tags, positive-length intervals, chronological order, and temporal bounds. A soft range term encourages the target length:

$$\begin{aligned} R_{\text{fmt}} &= \tfrac{1}{4}(r_{\text{tag}} + r_{\text{time}} + r_{\text{order}} + r_{\text{bound}}), \\ R_{\text{struct}} &= \lambda_{\text{fmt}} R_{\text{fmt}} + (1 - \lambda_{\text{fmt}}) r_{\text{range}}(n). \end{aligned} \tag{4}$$

Here $r_{\text{tag}}, r_{\text{time}}, r_{\text{order}}, r_{\text{bound}} \in \{0, 1\}$ are the four deterministic checks, n is the number of parsed events, and $\lambda_{\text{fmt}} = 0.7$. The range function uses $r_{\text{range}}(n) = 1$ for $2 \le n \le 5$ and decreases linearly with slope $\kappa = 0.5$ outside that interval, clipped at zero. Structure rewards parseability, not semantic truth or usefulness.

Pass 2 receives the original video, candidates, and a serialized, read-only copy of the sampled chain:

$$(\mathcal{R}^{(i)}, \hat{y}^{(i)}) \sim \pi_\theta(\mathcal{R}, \hat{y} \mid V, \mathcal{E}^{(i)}, \mathcal{O}, p_{\text{pred}}). \tag{5}$$

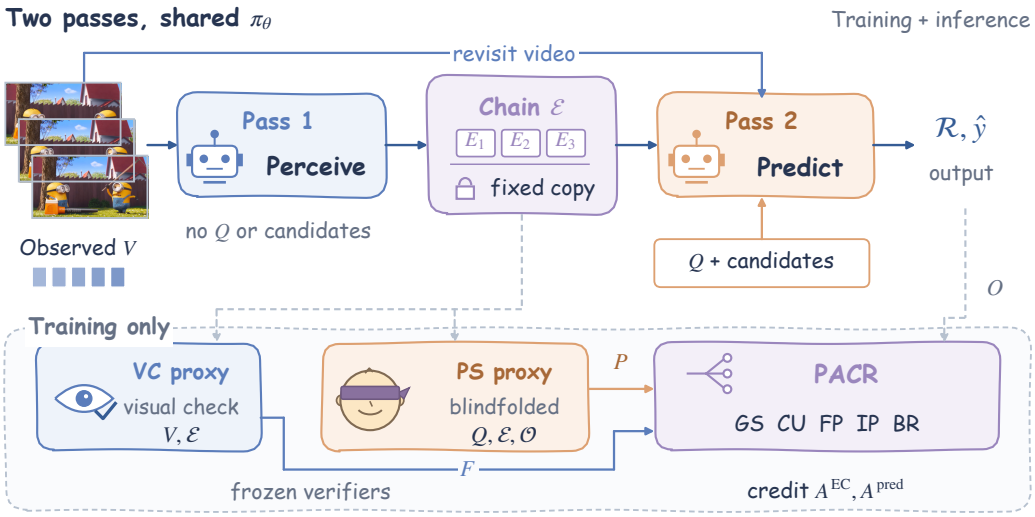


Figure 2: Overview of PGE-RL. The shared policy first samples an option-blind chain, then predicts from the video, task question, candidates, and a fixed copy of that chain. Pass 2 can revisit the video but cannot revise the chain. Frozen VC/PS proxies and five-state PACR provide training-time credit; the dashed branch is removed at inference. Appendix Figures A.1–A.3 provide the detailed view (Appendix A.1).

Pass 2 may re-examine V, but emits only reasoning and an answer label. It cannot modify the chain; event-like text is never parsed back into it. The copied chain is context, excluded from the generation loss. The policy uses 32 frames on FutureBench and 16 on VEPBench; AVEP uses its finite target-label vocabulary as candidates. Appendix A.1 provides an illustrative rollout; Appendix A.3 gives the documented prompts. The joint rollout is $\tau_i = (\mathcal{E}^{(i)}, \mathcal{R}^{(i)}, \hat{y}^{(i)})$. Although Pass 1 is option-blind at generation time, its sampled tokens still receive candidate-conditioned training feedback through predictive sufficiency. Option blindness therefore reduces direct copying without discarding the task signal needed to learn decision-relevant perception.

3.3 Dual Chain Verification

Timestamp-conditioned visual consistency. For event E_j, a frozen video-text encoder, building on contrastive vision–language pretraining (Radford et al., 2021), compares D_j with the claimed clip $V_j = V[t_j^{\text{s}} : t_j^{\text{e}}]$. To reward temporal specificity rather than scene-level similarity, let $\mathcal{N}_j$ contain non-overlapping intervals from the same video whose durations match that of V_j. We uniformly sample eight frames from each interval and use the frozen `alibaba-pai/VideoCLIP-XL` encoder in BF16 with gradients disabled. We use the contrastive interval score

$$\begin{aligned} \widetilde{s}_j &= s(D_j, V_j) - \max_{V' \in \mathcal{N}_j} s(D_j, V'), \\ a_j &= \text{clip}\left(\frac{\widetilde{s}_j - q_{\text{lo}}}{q_{\text{hi}} - q_{\text{lo}}}, 0, 1 \right), \\ R_{\text{align}} &= \frac{1}{n} \sum_{j=1}^{n} a_j. \end{aligned} \tag{6}$$

Invalid intervals receive zero alignment. Calibration is performed once per benchmark-specific run and then frozen: the initialized policy generates chains for 500 fixed training examples, and the claimed intervals together with duration-matched, temporally shifted controls determine $q_{\text{lo}} = Q_{0.05}$ and $q_{\text{hi}} = Q_{0.95}$. The mean supplies smooth training feedback; diagnostics use the minimum event score so one strong event cannot hide an unsupported one. No test example is used for calibration, and gradients never update the verifier.

Blindfolded predictive sufficiency. Visual alignment can favor generic descriptions. We therefore use the frozen text reasoner `Qwen/Qwen2.5-14B-Instruct`, following MoCA (Wang et al., 2026), that sees only $(Q, \mathcal{E}, \mathcal{O})$. It uses constrained scoring over the candidate labels and INSUFFICIENT; it must select a candidate from explicit chain evidence or return INSUFFICIENT; the video, policy reasoning, policy answer, and label are hidden. To prevent label-position shortcuts, candidate order is independently randomized for the policy and verifier, and both outputs are mapped back to semantic candidate identities before reward computation. To control option-only priors, define

$$\begin{aligned} \Delta_\phi &= \left[\log p_\phi(y^* \mid \mathcal{E}, \mathcal{O}) - \log p_\phi(y^* \mid \mathcal{O})\right]_+, \\ R_{\text{pred}} &= \mathbb{I}[g_\phi(\mathcal{E}, \mathcal{O}) = y^*]\ \text{clip}\big(\Delta_\phi / q_{\text{hi}}^{\Delta}, 0, 1\big). \end{aligned} \tag{7}$$

The scale q_{hi}^{Δ} is the 95th percentile of positive gains on the same training-only calibration set and is frozen thereafter. R_{pred} is a verifier-relative proxy for the marginal discriminative evidence contributed by the chain; it does not establish whether that evidence is factually correct and is therefore paired with R_{align}. The options-only subtraction reduces reward for candidate priors. Requiring a correct choice and positive gain still does not eliminate verifier errors; INSUFFICIENT permits, but cannot guarantee, appropriate abstention.

3.4 PERCEPTION-AWARE POLICY OPTIMIZATION

We combine the structural stabilizer with answer, alignment, and sufficiency rewards, using semantic weights that sum to one:

$$\begin{aligned} R_i = \eta(t) R_{\text{struct}}^{(i)} \\ + (1 - \eta(t))\big(\bar{\alpha} R_{\text{ans}}^{(i)} + \bar{\beta} R_{\text{align}}^{(i)} + \bar{\gamma} R_{\text{pred}}^{(i)}\big), \end{aligned} \tag{8}$$

where $\bar{\alpha} + \bar{\beta} + \bar{\gamma} = 1$. The structural weight anneals from η_0 to $\eta_{\min}$ early in training:

$$\eta(t) = \eta_{\min} + (\eta_0 - \eta_{\min}) \max\left(0, 1 - \frac{t}{T_{\text{anneal}}}\right). \tag{9}$$

We use $(\bar{\alpha}, \bar{\beta}, \bar{\gamma}) = (0.50, 0.25, 0.25)$, $\eta_0 = 0.20$, $\eta_{\min} = 0.05$, and $T_{\text{anneal}} = 0.2T$, where T is the total number of updates. For G rollouts of one example, GRPO standardizes this score:

$$A_i = \frac{R_i - \text{mean}(\{R_k\}_{k=1}^{G})}{\text{std}(\{R_k\}_{k=1}^{G}) + \epsilon_{\text{norm}}}. \tag{10}$$

Here $\epsilon_{\text{norm}} > 0$ stabilizes division. Broadcasting A_i to both passes can punish a useful history when downstream reasoning fails or reinforce a hallucinated history after a lucky answer. We therefore route credit according to the independently measured outcome, sufficiency, and visual reliability of the same rollout.

We derive outcome $O_i = R_{\text{ans}}^{(i)}$, sufficiency $P_i = \mathbb{I}[R_{\text{pred}}^{(i)} \geq \tau_P]$, and conservative visual reliability $F_i = R_{\text{fmt}}^{\text{hard},(i)} \min_j a_j^{(i)}$, where $R_{\text{fmt}}^{\text{hard}} = \prod_{k \in \{\text{tag,time,order,bound}\}} r_k$ for a nonempty parsed chain and is zero otherwise. Thus F_i is zero for an empty or malformed chain and otherwise is limited by its weakest event alignment. They identify five mutually exclusive states: grounded success GS when $O_i = P_i = 1, F_i \geq \tau_F$; correct but ungrounded CU when $O_i = 1$ and either check fails; false perception FP when $O_i = 0, F_i < \tau_F$; incomplete perception IP when $O_i = 0, F_i \geq \tau_F, P_i = 0$; and bad downstream reasoning BR when $O_i = 0, F_i \geq \tau_F, P_i = 1$. These labels are diagnostic routing states, not verified causes of errors. Chain-token credit is

$$A_i^{\text{EC}} = \begin{cases} A_i, & \text{GS}, \\ 0, & \text{CU}, \\ A_i - \rho_f (1 - F_i)|A_i|, & \text{FP}, \\ A_i - \rho_m F_i |A_i|, & \text{IP}, \\ A_i + \rho_p F_i |A_i|, & \text{BR}, \end{cases} \tag{11}$$

with $0 \leq \rho_m < \rho_f \leq 1$. The larger coefficient for FP emphasizes unsupported assertions; the realized adjustment also depends on F_i and $|A_i|$. In BR, the adjustment attenuates negative

chain advantages and increases positive ones. We set $(\tau_F, \tau_P) = (0.60, 0.50)$ and $(\rho_f, \rho_m, \rho_p) = (0.20, 0.10, 0.20)$. The minimum-event definition of F_i is intentionally conservative: a single unsupported event can invalidate a premise even when the rest of the chain is well aligned. Prediction-token credit is

$$A_i^{\text{pred}} = \begin{cases} \zeta[A_i]_+, & \mathsf{CU}, \\ A_i, & \text{otherwise}, \end{cases} \qquad 0 < \zeta < 1, \tag{12}$$

so a correct answer that fails either proxy receives attenuated positive prediction credit and no chain credit; we use $\zeta = 0.20$. If $m_{i,t}^{\text{EC}}$ and $m_{i,t}^{\text{pred}}$ denote generated tokens from the two passes, their routed advantage is

$$A_{i,t} = m_{i,t}^{\text{EC}} A_i^{\text{EC}} + m_{i,t}^{\text{pred}} A_i^{\text{pred}}. \tag{13}$$

We insert $A_{i,t}$ into the clipped GRPO objective (Appendix A.2), with clipping parameter $\epsilon_{\text{clip}} = 0.2$, and KL penalty $\lambda_{\text{KL}} = 0.04$ to the frozen initial policy. Prompt tokens, visual tokens, padding, and the read-only serialized chain are excluded from the generation masks. The Pass-1 and Pass-2 token losses are averaged separately before being averaged together, so sequence length does not determine stage weighting. Pass-1 tokens are optimized through their own sampled log probabilities; copied Pass-2 chain tokens carry no generation-loss term. Training settings and the full objective are in Appendix A.2.

3.5 Training and Inference

PGE-RL starts directly from the instruction-tuned `Qwen/Qwen3-VL-8B-Instruct`; frozen verifiers and deterministic structure checks make event-chain annotations and an SFT cold start unnecessary. FutureBench, VEPBench, and AVEP are trained in separate benchmark-specific runs, so neither reward calibration nor policy updates mix benchmark examples. At inference, the policy performs the same option-blind chain generation followed by video-aware prediction (or constrained scoring over the finite AVEP target vocabulary). Verifiers, rewards, group sampling, and all diagnostic routing logic are removed. Checkpoints are selected by validation answer accuracy, and the final reported policy uses the selected checkpoint without test-time reward computation.

4 Experiments

4.1 Setup

Evaluation scope. Tables preserve the supplied point estimates. Some FutureBench AVG entries cannot be reconstructed using the official 1,056-example category weights; VEPBench count denominators and AVEP P/R/F1 aggregation also need reconciliation with per-example records (Appendix A.6). Rankings are descriptive, not statistically validated; no seeds or uncertainty estimates are available.

Benchmarks and metrics. We evaluate on FutureBench (Wang et al., 2025a), VEPBench, and AVEP (Su et al., 2025). FutureBench reports 1/2/3-Hop accuracy, interpolation, and AVG. Its questions provide outcome or intermediate-event anchors; withholding future video frames does not imply the absence of such textual context. VEPBench contains 3,000 verified examples with five semantic option types: the recorded future (type A), a plausible but unrecorded future (type B), and three incorrect alternatives (types C–E). Display order is randomized for every instance; the selected display label is mapped back to its semantic type before computing SR@A, SR@AB, and SR@C/D/E. We use its source-disjoint 17K-video training set. We report closed-set evaluation only. Predicted option labels are parsed and mapped to their canonical semantic types; Appendix A.3 provides the benchmark evaluation prompt and candidate-order randomization protocol.

AVEP evaluates structured verb, noun, and action predictions. The main paper reports verb scores and noun/action precision, recall, and F1 on test and validation splits; Appendix A.4 expands the baseline inventory. Each target uses a finite label vocabulary; the benchmark-specific prediction interface is described with the AVEP results below.

Implementation details. Vanilla baselines use the benchmark prompt without event-chain scaffolding; model and method references appear beside the table entries. PGE-RL starts from

`Qwen/Qwen3-VL-8B-Instruct` without event-chain SFT. Separate FutureBench/VEPBench runs use 32/16 uniformly sampled frames and eight A100-80GB GPUs. The reported RL configuration uses 300 updates, $B = 8$, $G = 8$, and 19,200 sampled trajectories; PGE-RL allows 256/384 generated tokens for Passes 1/2. Learning rates are 5×10^{-7} for the language model/projector and 1×10^{-7} for the vision encoder, with BF16 and ZeRO-3. SFT uses gold answers without teacher chains; answer-only GRPO removes semantic chain rewards. These one-pass controls share the stated data, update, frame, and total token allowances, but are not interface- or measured-compute-matched to the two-pass method. Checkpoints are selected by validation answer accuracy. Appendix A.6 distinguishes available results from unmeasured controls and records missing reproduction details.

4.2 Main Results

Results on FutureBench. Table 1 compares PGE-RL on FutureBench (Wang et al., 2025a). It improves Qwen3-VL-8B on every split, raising reported AVG from 55.50 to 86.10, and leads the SFT and GRPO controls throughout. Gains over the controlled backbone are $27.00/27.30/34.90$ points on 1/2/3-Hop and 31.70 on interpolation, covering both local and longer-range dependencies.

Table 1: FutureBench results (%). Bold/underline mark best/second best; PGE-RL is ours. AVG aggregation requires reconciliation (Section 4.1).

Model	Method	Frames	FutureBench↑				
			1-Hop	2-Hop	3-Hop	Interp.	AVG
Commercial Models							
GPT-5.4 (OpenAI, 2026)	Vanilla	32	59.80	57.70	62.10	55.30	57.82
Gemini 3.1 Pro Preview (Google, 2026)	Vanilla	32	61.00	59.00	63.20	56.50	58.97
Open-Source Models							
Qwen2.5-VL-3B (Bai et al., 2025b)	Vanilla	32	44.90	50.80	44.50	46.20	45.80
Qwen2.5-VL-7B (Bai et al., 2025b)	Vanilla	32	57.70	56.60	50.00	50.70	53.14
Qwen2.5-VL-32B (Bai et al., 2025b)	Vanilla	32	66.60	63.00	63.70	54.80	59.74
Qwen2.5-VL-72B (Bai et al., 2025b)	Vanilla	32	55.20	68.30	63.80	53.50	58.83
Qwen3-VL-4B (Bai et al., 2025a)	Vanilla	32	46.50	52.50	46.50	48.50	48.00
Qwen3-VL-8B (Bai et al., 2025a)	Vanilla	32	60.00	58.50	52.50	52.50	55.50
GLM-4.1V-9B (GLM-V Team, 2025)	Vanilla	32	30.20	42.40	51.80	47.10	44.38
InternVL3-8B (Zhu et al., 2025)	Vanilla	32	53.90	57.80	63.20	54.60	57.12
Kimi-VL-A3B (Kimi Team, 2025)	Vanilla	32	44.20	42.90	51.60	52.40	48.47
LLaVA-NeXT-Video-7B (Zhang et al., 2024)	Vanilla	32	48.80	49.50	40.40	43.90	44.94
MiMo-VL-7B (Xiaomi LLM-Core Team, 2025)	Vanilla	32	59.40	59.10	50.20	43.70	50.55
Visual Reasoning							
Video-CoE (7B) (Su et al., 2026)	RL	32	<u>80.90</u>	<u>83.90</u>	<u>71.60</u>	<u>71.40</u>	<u>75.00</u>
Video-R1 (7B) (Feng et al., 2025)	RL	32	71.80	69.50	65.40	65.99	67.47
VideoChat-R1 (7B) (Li et al., 2025b)	RL	32	48.50	47.80	44.20	46.42	46.59
LVR (7B) (Li et al., 2025a)	RL	32	22.5	26.4	22.9	17.6	21.0
Monet (7B) (Wang et al., 2025b)	RL	32	46.8	47.2	45.3	49.7	47.9
SwimBird (8B) (Tong et al., 2026)	RL	32	59.0	66.8	64.7	61.8	62.8
Qwen3-VL-8B (Bai et al., 2025a)	SFT	32	53.22	56.39	57.71	47.19	53.65
Qwen3-VL-8B	GRPO	32	62.81	56.48	52.44	51.32	53.67
PGE-RL (8B) (ours)	RL	32	**87.00**	**85.80**	**87.40**	**84.20**	**86.10**

Results on VEPBench. Table 2 gives the 16-frame comparison. PGE-RL reaches 68.95 SR@A and 74.90 SR@AB, improving the vanilla backbone by $29.02/14.23$ points. It achieves the lowest SR@E at 0.61, though it does not lead SR@C or SR@D. The gap $\mathrm{SR@AB} - \mathrm{SR@A}$ measures how often a model selects the plausible but unrecorded type-B future rather than the recorded type-A future. PGE-RL reduces this gap from 20.74 for vanilla Qwen3-VL-8B to 5.95, a 14.79-point reduction. This indicates fewer type-B choices in the reported evaluation; it does not by itself isolate visual evidence use from language priors.

Table 2: VEPBench results (%). Asterisks mark primary metrics; bold/underline denote best/second best.

Model	Method	Frames	SR@A↑*	SR@AB↑*	SR@C↓	SR@D↓	SR@E↓
Commercial Models							
GPT-5.4 (OpenAI, 2026)	Vanilla	16	52.06	63.43	30.76	**4.00**	1.80
Gemini 3.1 Pro Preview (Google, 2026)	Vanilla	16	51.79	58.65	34.76	5.93	0.65
Open-Source Models							
Qwen2.5-VL-3B (Bai et al., 2025b)	Vanilla	16	36.65	57.53	24.00	14.78	3.70
Qwen2.5-VL-7B (Bai et al., 2025b)	Vanilla	16	41.38	65.75	27.94	4.80	1.50
Qwen2.5-VL-32B (Bai et al., 2025b)	Vanilla	16	54.48	72.47	20.03	5.00	2.46
Qwen3-VL-8B (Bai et al., 2025a)	Vanilla	16	39.93	60.67	32.07	5.93	1.33
Qwen3-VL-32B (Bai et al., 2025a)	Vanilla	16	53.80	63.94	31.40	<u>4.02</u>	<u>0.64</u>
GLM-4.6V-Flash-9B (Z.ai, 2026)	Vanilla	16	51.40	59.78	**7.26**	9.50	23.46
InternVL3-8B (Zhu et al., 2025)	Vanilla	16	42.88	68.79	21.71	7.44	2.07
Keye-VL-8B (Kwai Keye Team, 2025)	Vanilla	16	38.77	60.40	24.07	10.87	4.67
Kimi-VL-A3B (Kimi Team, 2025)	Vanilla	16	46.48	<u>74.38</u>	15.81	6.70	3.11
LLaVA-NeXT-Video-7B (Zhang et al., 2024)	Vanilla	16	52.97	70.33	15.83	8.40	5.43
MiMo-VL-7B (Xiaomi LLM-Core Team, 2025)	Vanilla	16	36.83	67.47	24.53	5.60	2.40
Visual Reasoning							
Video-Chat-A1 (7B) (Wang et al., 2025c)	RL	16	41.17	64.17	26.50	7.10	2.23
Video-R1 (7B) (Feng et al., 2025)	RL	16	47.93	71.53	19.80	7.17	1.50
LVR (7B) (Li et al., 2025a)	RL	16	55.41	68.40	<u>11.17</u>	12.53	7.90
Monet (7B) (Wang et al., 2025b)	RL	16	53.41	70.41	<u>11.17</u>	10.53	7.89
SwimBird (8B) (Tong et al., 2026)	RL	16	<u>56.41</u>	68.41	13.17	10.52	7.90
Qwen3-VL-8B (Bai et al., 2025a)	SFT	16	51.37	64.54	14.07	13.60	7.80
Qwen3-VL-8B	GRPO	16	54.41	69.41	12.17	11.53	6.90
PGE-RL (8B) (ours)	RL	16	**68.95**	**74.90**	14.10	10.39	**0.61**

Results on AVEP. Table 3 reports AVEP (Su et al., 2025) results under separate benchmark-specific training, not zero-shot transfer. Pass 2 scores the finite verb, noun, or action label vocabulary by constrained decoding. PGE-RL exceeds Qwen3-VL-8B-GRPO on all reported metrics. Against CoE-GRPO-7B (Su et al., 2026), it improves noun and action F1 on both splits; CoE retains higher validation verb and action recall scores. Appendix Tables A.3 and A.4 expand the baseline inventory. All comparisons use reported point estimates.

Table 3: Structured prediction on AVEP (%). Each cell is test/validation. Higher is better; bold/underline mark best/second-best among these methods.

Method	Verb (Test/Val)↑	Noun (Test / Val)↑			Action (Test / Val)↑		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
Qwen3-VL-8B-GRPO	10.00/11.00	48.00/47.80	91.00/92.00	63.00/63.80	5.20/6.00	9.50/10.00	6.70/7.50
CoE-GRPO-7B	<u>12.24</u>/**18.75**	<u>49.88</u>/<u>48.68</u>	<u>93.95</u>/<u>93.54</u>	<u>65.16</u>/<u>64.03</u>	<u>6.62</u>/<u>7.70</u>	<u>11.09</u>/**13.80**	<u>8.29</u>/<u>9.88</u>
PGE-RL (8B) (ours)	**15.00**/<u>16.00</u>	**52.00/51.00**	**95.00/96.00**	**68.00/69.00**	**7.50/8.50**	**12.00**/<u>13.00</u>	**9.00/10.00**

4.3 Ablation Study

Component ablations. Table 4 tests the joint objectives and routing. VC alone favors alignment and PS alone sufficiency (Figure 3(a)); their combination balances both proxies.

Removing two-pass generation or PACR reduces GSR from 70.1% to 53.7% or 62.3%. Align. requires the weakest interval to pass $F_i \geq \tau_F$, Suff. requires $P_i = 1$, and GSR also requires a correct answer. These verifier-relative diagnostics cover both benchmarks, but their populations and aggregation weights are unrecorded.

A 500-example audit with three independent annotators found more than 90% majority agreement for each diagnostic; classwise statistics and inter-annotator agreement are unavailable, so this does not establish calibration or evaluator transfer.

Table 4: Component ablations (%). FB: AVG; VEP: SR@A. Higher is better; diagnostics are verifier-relative and N/A means no chain.

Variant	FB	VEP	Align.	Suff.	GSR
GRPO only	53.67	54.41	N/A	N/A	N/A
VC only	62.40	51.20	**85.4**	48.6	41.2
PS only	70.10	60.50	52.7	**86.3**	44.5
No two-pass	65.30	55.80	74.1	72.4	53.7
No PACR	72.80	63.40	78.5	80.1	62.3
Full	**86.10**	**68.95**	82.3	84.7	**70.1**

Behavioral chain dependence. With policy and decoding fixed, chain removal and corruption lower scores (Table 5; Figure 3(b)). Full uses video, chain, and options; removing the chain while retaining video reduces FB/VEP by 12.60/5.15 points, while swapping it reduces them by 24.90/18.45. Flip measures sensitivity with unreported pooling weights, not answer quality. These interventions support behavioral dependence but do not identify causality because corruption can introduce distribution shift.

Table 5: Chain-input interventions. Score (%) and drop (pp) are FB AVG / VEP SR@A. Flip is prediction change (%).

Evidence	Score FB / VEP	Drop FB / VEP	Flip
Options only	58.20 / 43.70	27.90 / 25.25	∼74
No chain	73.50 / 63.80	12.60 / 5.15	38
Chain only	70.20 / 61.80	15.90 / 7.15	48
Generic	68.40 / 57.20	17.70 / 11.75	52
Corrupt time	65.80 / 54.60	20.30 / 14.35	63
Swapped	61.20 / 50.50	24.90 / 18.45	72
Full	**86.10 / 68.95**	0.00 / 0.00	0

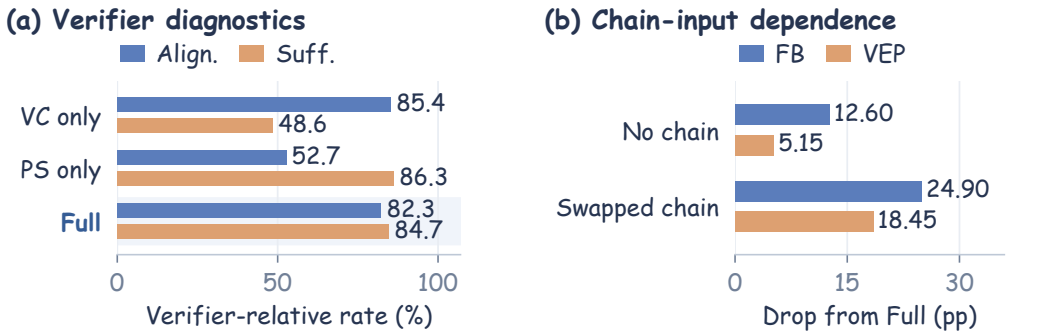


Figure 3: Verifier-relative diagnostics and chain-intervention drops. Point estimates require population and aggregation reconciliation; no uncertainty estimates are available.

5 Discussion and Conclusion

PGE-RL combines immutable, option-blind histories, video-aware prediction, and five-state routing. The recorded estimates favor the combination under benchmark-specific training; interventions support behavioral chain dependence.

VC/PS remain verifier-relative proxies, and PS does not establish factual truth. The human audit lacks classwise statistics. Evaluator transfer, verifier-backbone and routing-sensitivity results are unavailable; one-pass baselines do not isolate interface effects. Corrupted inputs may induce distribution shift. Unreconciled metric aggregation and absent uncertainty estimates limit the quantitative conclusions. These results therefore do not establish statistical significance, factual-grounding guarantees, or cross-benchmark transfer.

AI USE STATEMENT

Generative AI assisted language editing and coding. During preparation and revision of this manuscript, AI also assisted structural suggestions, LaTeX template migration and formatting, clarification of methodological exposition, scientific figure preparation, literature and policy checks, and critical interpretation of the reported results. The language models that are components of the research method are described in Section 3. This statement records the confirmed uses; it does not assert that AI-assisted content has been independently validated. Responsibility for the final text, claims, references, code, and results rests with the authors.

ETHICS STATEMENT

Future-event prediction can misrepresent an observed person's actions or intentions. Verifier agreement and a correct candidate choice do not establish that a textual history is factual, nor do these benchmark evaluations establish suitability for decisions about individuals. Video reuse and release require respect for dataset terms, privacy, and applicable model licenses. Human assessment is reported in Appendix A.5; detailed annotation and governance records are not included in the supplied materials.

REPRODUCIBILITY STATEMENT

Section 3 specifies the information boundaries, proxy rewards, and routed advantages. Appendix Sections A.2–A.6 provide documented prompts, optimization equations, training settings, full AVEP results, and available chain interventions. It also identifies missing evaluation records and implementation details. No executable training/evaluation package, per-example predictions, or exact trained checkpoint revisions accompanies these PDFs; the recorded point estimates have not been independently reproduced here.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report, 2025a.

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, 2025b.

Zixu Cheng, Da Li, Jian Hu, Yuhang Zang, Ziquan Liu, Shaogang Gong, and Wei Li. Graph-Thinker: Reinforcing temporally grounded video reasoning with event graph thinking. *arXiv preprint arXiv:2602.17555*, 2026. doi: 10.48550/arXiv.2602.17555.

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, 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*, 2025.

GLM-V Team. GLM-4.1V-Thinking and GLM-4.5V: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025.

Google. Gemini 3.1 Pro Preview model documentation. `https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview`, 2026. Accessed July 29, 2026.

Chenhui Gou, Ziyu Ma, Zicheng Duan, Haoyu He, Feng Chen, Akide Liu, Bohan Zhuang, Jianfei Cai, and Hamid Rezatofighi. An empirical study on how Video-LLMs answer video questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18587–18597, 2026.

Tianxiang Jiang, Linquan Wu, Sheng Xia, Songze Li, Ziang Yan, Haoyu Yang, Yu Qiao, and Yi Wang. Imagine before you predict: Interleaved latent visual reasoning for video event prediction. *arXiv preprint arXiv:2606.05769*, 2026a. doi: 10.48550/arXiv.2606.05769.

Yifan Jiang, Yueying Wang, Rui Zhao, Toufiq Parag, Zhimin Chen, Zhenyu Liao, and Jayakrishnan Unnikrishnan. VideoP2R: Video understanding from perception to reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 8303–8313, 2026b.

Kimi Team. Kimi-VL technical report, 2025.

Kwai Keye Team. Kwai Keye-VL technical report, 2025.

Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 8769–8784. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.706.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning, 2025a.

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, 2025b.

Baoyu Liang, Qile Su, Shoutai Zhu, Yuchen Liang, and Chao Tong. VidEvent: A large dataset for understanding dynamic evolution of events in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5128–5136, 2025.

Junzhang Liu, Zhecan Wang, Hammad Ayyubi, Haoxuan You, Chris Thomas, Rui Sun, Shih-Fu Chang, and Kai-Wei Chang. Detecting multimodal situations with insufficient context and abstaining from baseless predictions, 2024.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-RFT: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

OpenAI. GPT-5.4 model documentation. `https://developers.openai.com/api/docs/models/gpt-5.4`, 2026. Accessed July 29, 2026.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. VisualCOMET: Reasoning about the dynamic context of a still image. In *Proceedings of the European Conference on Computer Vision*, pp. 508–524, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J. Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024.

Chufan Shi, Cheng Yang, Yaokang Wu, Linhao Jin, Bo Shui, Taylor Berg-Kirkpatrick, and Xuezhe Ma. Are VLMs seeing or just saying? uncovering the illusion of visual re-examination. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Qile Su, Shoutai Zhu, Shuai Zhang, Baoyu Liang, and Chao Tong. EventFormer: A node-graph hierarchical attention transformer for action-centric video event prediction. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4698–4707, 2025. doi: 10.1145/3746027.3755556.

Qile Su, Jing Tang, Rui Chen, Lei Sun, and Xiangxiang Chu. Video-CoE: Reinforcing video event prediction via chain of events, 2026.

Jintao Tong, Shilin Yan, Hongwei Xue, Xiaojun Tang, Kunyu Shi, Guannan Zhang, Ruixuan Li, and Yixiong Zou. SwimBird: Eliciting switchable reasoning mode in hybrid autoregressive MLLMs, 2026.

Haonan Wang, Hongfu Liu, Xiangyan Liu, Chao Du, Kenji Kawaguchi, Ye Wang, and Tianyu Pang. Fostering video reasoning via next-event prediction, 2025a.

Haozhe Wang, Qixin Xu, Changpeng Wang, Taofeng Xue, Chong Peng, Wenhu Chen, and Fangzhen Lin. Bad seeing or bad thinking? rewarding perception for multimodal reasoning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond images and language, 2025b.

Zikang Wang, Boyu Chen, Zhengrong Yue, Yi Wang, Yu Qiao, Limin Wang, and Yali Wang. VideoChat-A1: Thinking with long videos by chain-of-shot reasoning, 2025c.

Yana Wei, Liang Zhao, Kangheng Lin, En Yu, Yuang Peng, Runpei Dong, Jianjian Sun, Haoran Wei, Zheng Ge, Xiangyu Zhang, and Vishal M. Patel. Perception in reflection. *arXiv preprint arXiv:2504.07165*, 2025.

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.

Xiaomi LLM-Core Team. MiMo-VL technical report, 2025.

Z.ai. GLM-4.6V-Flash-9B model card. Hugging Face model card, 2026. URL `https://huggingface.co/zai-org/GLM-4.6V-Flash`.

Chi Zhang, Haibo Qiu, Qiming Zhang, Yufei Xu, Zhixiong Zeng, Siqi Yang, Peng Shi, Lin Ma, and Jing Zhang. Perceptual-evidence anchored reinforced learning for multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41111–41120, 2026.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-VL: Learning to reason with multimodal large language models via step-wise group relative policy optimization, 2025a.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: A strong zero-shot video understanding model. LLaVA technical report, 2024. URL `https://llava-vl.github.io/blog/2024-04-30-llava-next-video/`.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data, 2025b.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025.

A Appendix

Organization. Section A.1 details the interfaces and worked examples; Section A.2 gives optimization, configuration, and control settings; Section A.3 documents prompts and evaluation; Section A.4 presents additional results and interventions; Section A.5 discusses human assessment and unmeasured reliability tests; and Section A.6 records evidence and reproducibility limits.

A.1 Method Interfaces and Illustrative Examples

The detailed framework expands the policy, verifier, and routing panels of main-paper Figure 2. Video frames illustrate the interface rather than document a measured rollout.

A.1.1 Two-pass policy interface

Figure A.1 separates candidate-blind chain sampling from prediction with the immutable chain. The video remains available to the second pass.

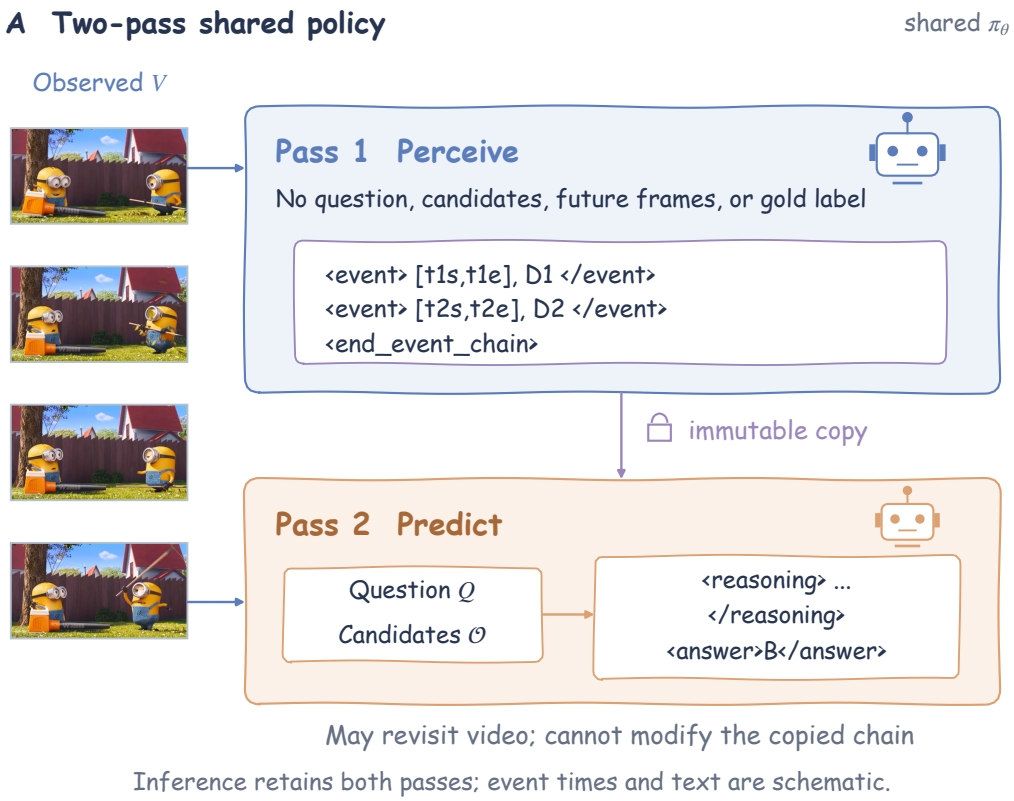


Figure A.1: Detailed framework, panel A: the two-pass policy interface. The event chain is generated before candidates are exposed and copied unchanged into the prediction pass. Pass 2 also sees the video, task question, and candidates; it can revisit the video but cannot rewrite the chain. Both passes remain at inference.

A.1.2 Frozen training-time verifiers

Figure A.2 distinguishes visual consistency from predictive sufficiency. Both are verifier-relative proxies; the blindfolded sufficiency check does not verify factual truth.

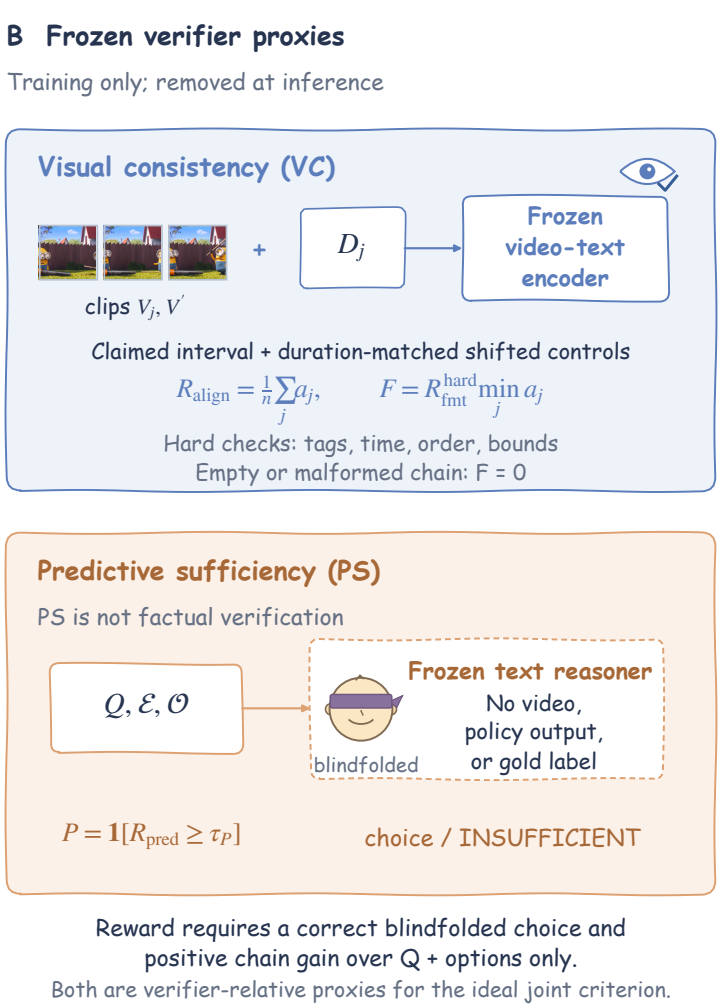


Figure A.2: Detailed framework, panel B: frozen training-time verifiers. VC estimates support for each description in its claimed interval. The blindfolded PS verifier sees the task question, chain, and candidates but no video or policy answer. It measures predictive usefulness rather than factual truth. These scores approximate the ideal VC/PS criteria; the verifiers are removed at inference.

A.1.3 Structural Reward Details

Let n be the number of parsed events. The soft range reward used by the main-paper structural objective is

$$r_{\text{range}}(n) = \begin{cases} 1, & 2 \leq n \leq 5, \\ \max(0, 1 - \kappa(2 - n)), & n < 2, \\ \max(0, 1 - \kappa(n - 5)), & n > 5, \end{cases} \tag{A.1}$$

where $\kappa = 0.5$ controls the slope outside the valid range; the main structural reward uses $\lambda_{\text{fmt}} = 0.7$. The four deterministic format terms check matched event and termination tags (r_{tag}), parsable intervals satisfying $t_j^{\text{s}} < t_j^{\text{e}}$ (r_{time}), chronological ordering without pathological overlap (r_{order}), and timestamps within the observed duration (r_{bound}). Invalid events fail the corresponding check.

For routing, all four checks must pass on a nonempty chain:

$$R_{\text{fmt}}^{\text{hard}} = \begin{cases} \prod_{k \in \{\text{tag,time,order,bound}\}} r_k, & n > 0, \\ 0, & n = 0. \end{cases} \tag{A.2}$$

The reliability score is $F_i = R_{\text{fmt}}^{\text{hard},(i)} \min_j a_j^{(i)}$ for a nonempty parsed chain and $F_i = 0$ for an empty chain. The hard gate is distinct from the soft average R_{fmt} and the event-count preference: passing it establishes a valid interface, not factual accuracy.

A.1.4 Perception-aware credit routing

Figure A.3 summarizes the five diagnostic states and the resulting chain and prediction advantages. The training objective and its token masks are specified in Section A.2.

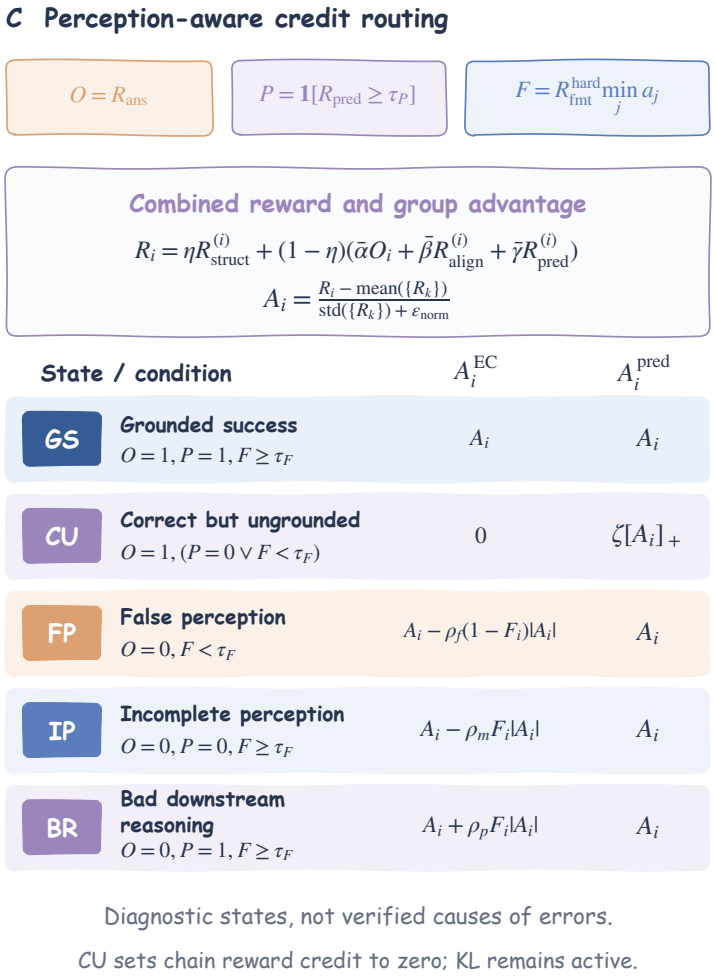


Figure A.3: Detailed framework, panel C: perception-aware credit routing. Outcome, hard-format-gated visual reliability, and predictive sufficiency determine five routing states and distinct advantages for chain and prediction tokens. The state names describe diagnostic routing rather than established causes of errors. CU suppresses chain reward credit; the KL regularizer still applies. Reward computation and routing are absent at inference.

A.1.5 Visual walkthrough

Figure A.4 illustrates how Pass 2 can reinspect the observed video while keeping the Pass-1 event chain fixed. The later frames are separated as excluded future context.

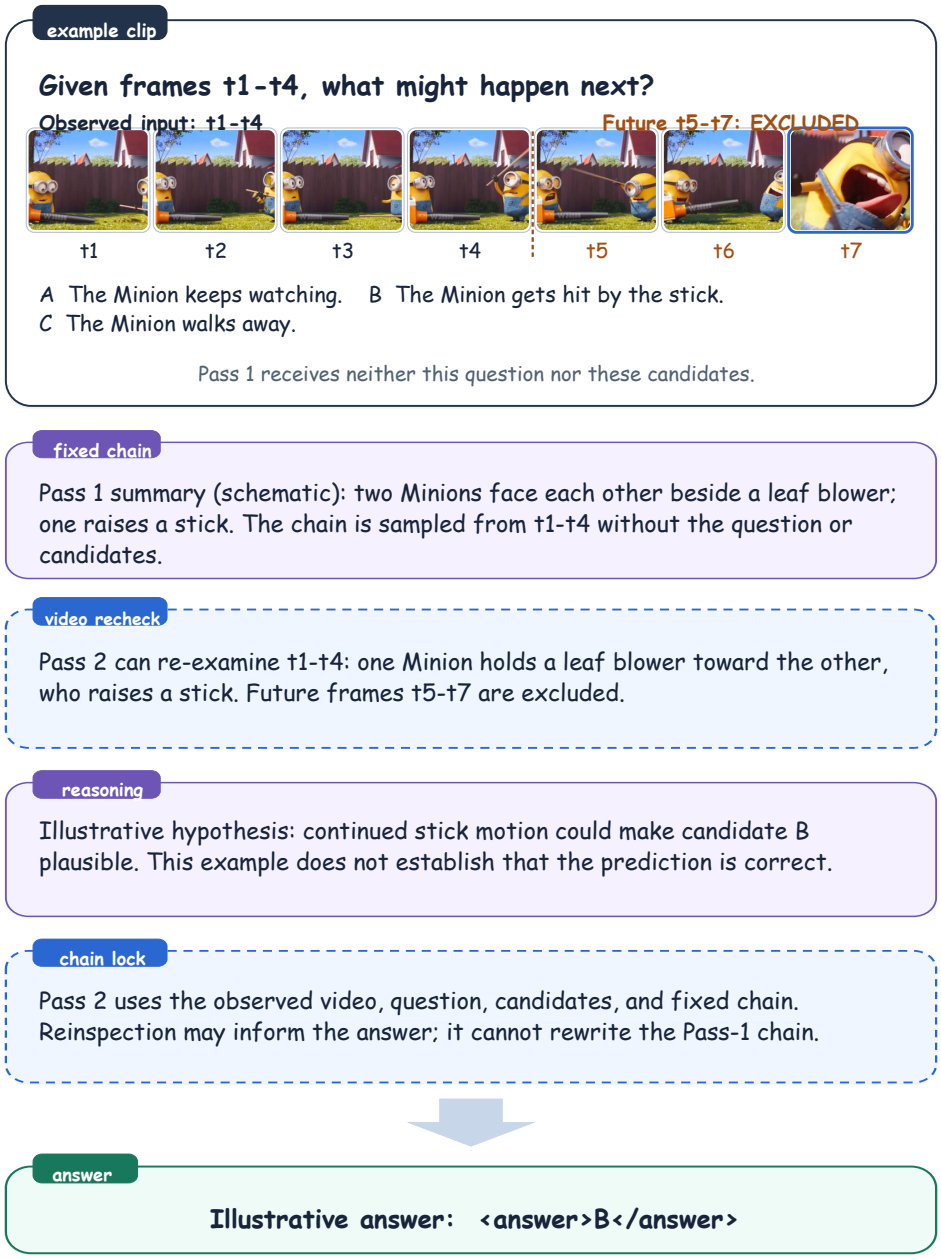


Figure A.4: Schematic illustration of Pass-2 video reinspection with a frozen event chain, adapted from the author-provided slide. Only frames t_1–t_4 are observed inputs; t_5–t_7 are shown separately as future context and are excluded from both passes. Pass 1 samples the chain without the question or candidates. Pass 2 receives the observed video, question, candidates, and immutable chain. The prose and candidate-B answer are illustrative, not logged model outputs or evidence of a correct prediction.

A.1.6 Illustrative Rollout and Parsed Representation

The serialization below is a separate schematic example of syntax and parsing; it does not transcribe Figure A.4. Figure A.5 illustrates the two-pass interface and the event tuples it represents. It is a schematic reconstruction, not a logged model output or a human-verified grounding example. In particular, the candidate list belongs only to Pass 2: Pass 1 does not receive it when sampling the history.

Pass 1 input: observed video and perception prompt.

Illustrative Pass 1 output:

```
<event>
Time:  0.0-2.1
Description:  Two Minions face each other.
</event>
<event>
Time:  2.1-4.3
Description:  A stick moves toward one Minion.
</event>
<event>
Time:  4.3-5.8
Description:  The gap rapidly closes.
</event>
<end_event_chain>
```

Parsed representation: $E_1 = ([0.0, 2.1], D_1)$, $E_2 = ([2.1, 4.3], D_2)$, $E_3 = ([4.3, 5.8], D_3)$, where each D_j is the corresponding description above.

Pass 2 input: observed video, the question "What is most likely to happen next?", the immutable parsed chain, and candidates A: Keeps watching; B: Gets hit by the stick; C: Walks away.

Illustrative Pass 2 output:

```
<reasoning>
The stick continues toward contact.  B best continues the final
observed state.
</reasoning>
<answer>B</answer>
```

Figure A.5: Schematic serialization and its parsed event tuples. The tag convention follows the documented prompts in Section A.3. Pass 2 can re-examine the video but cannot modify the serialized chain. This illustration supplies no empirical evidence about parser success, visual consistency, or prediction accuracy.

A.2 Optimization, Configuration, and Controls

A.2.1 Optimization and Calibration Details

Verifier calibration. The visual verifier is `alibaba-pai/VideoCLIP-XL`, evaluated in BF16 with gradients disabled and eight uniformly sampled frames per claimed interval. Before RL, the initialized policy generates chains for 500 fixed training examples from each separately trained benchmark. For event description D_j, claimed interval V_j, and the set $\mathcal{N}_j$ of non-overlapping, duration-matched intervals from the same video, calibration and training use

$$\begin{aligned} \widetilde{s}_j &= s(D_j, V_j) - \max_{V' \in \mathcal{N}_j} s(D_j, V'), \\ q_{\mathrm{lo}} &= Q_{0.05}(\{\widetilde{s}_j\}_{\mathrm{cal}}), \qquad q_{\mathrm{hi}} = Q_{0.95}(\{\widetilde{s}_j\}_{\mathrm{cal}}), \\ a_j &= \mathrm{clip}\left(\frac{\widetilde{s}_j - q_{\mathrm{lo}}}{q_{\mathrm{hi}} - q_{\mathrm{lo}}}, 0, 1\right). \end{aligned} \tag{A.3}$$

The calibration set contains training-only claimed intervals and their temporally shifted, non-overlapping controls. Quantiles are estimated independently for each benchmark-specific run and then frozen; invalid intervals receive $a_j = 0$. The frozen text verifier is `Qwen/Qwen2.5-14B-Instruct`; it uses constrained scoring over the available labels and Insufficient. Its predictive-gain scale q_{hi}^{Δ} is the 95th percentile of positive gains on the same training-only calibration set. All quantiles are estimated separately and frozen throughout training and evaluation.

Structural schedule and routed GRPO objective. The structural coefficient follows

$$\eta(t) = \eta_{\min} + (\eta_0 - \eta_{\min}) \max\left(0, 1 - \frac{t}{T_{\mathrm{anneal}}}\right). \tag{A.4}$$

Let $m_{i,t}^{\mathrm{EC}}$ and $m_{i,t}^{\mathrm{pred}}$ indicate generated Pass-1 and Pass-2 tokens. The routed advantage and importance ratio are

$$\begin{aligned} A_{i,t} &= m_{i,t}^{\mathrm{EC}} A_i^{\mathrm{EC}} + m_{i,t}^{\mathrm{pred}} A_i^{\mathrm{pred}}, \\ r_{i,t}(\theta) &= \frac{\pi_\theta(y_{i,t} \mid h_{i,t})}{\pi_{\theta_{\mathrm{old}}}(y_{i,t} \mid h_{i,t})}. \end{aligned} \tag{A.5}$$

For a Pass-1 token, $h_{i,t}$ contains the video and perception prompt; for a Pass-2 token, it contains the video, task question, fixed chain, candidates, and prediction prompt. With

$$\begin{aligned} \ell_{i,t}(\theta) = \; & \min(r_{i,t}A_{i,t}, \mathrm{clip}(r_{i,t}, 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}})A_{i,t}) \\ & - \lambda_{\mathrm{KL}} D_{\mathrm{KL}}(\pi_\theta(\cdot \mid h_{i,t}) \| \pi_{\mathrm{ref}}(\cdot \mid h_{i,t})), \end{aligned} \tag{A.6}$$

we separately normalize the two generated stages:

$$\begin{aligned} \mathcal{L} = -\frac{1}{2G} \sum_{i=1}^{G} \Bigg[& \frac{\sum_{t \in \mathcal{T}_i^{\mathrm{EC}}} \ell_{i,t}}{|\mathcal{T}_i^{\mathrm{EC}}|} \\ & + \frac{\sum_{t \in \mathcal{T}_i^{\mathrm{pred}}} \ell_{i,t}}{|\mathcal{T}_i^{\mathrm{pred}}|} \Bigg]. \end{aligned} \tag{A.7}$$

Prompt, visual, padding, and the read-only serialized chain in Pass 2 are excluded from the generation masks. The reference policy is the frozen initialization shared by both passes. The reward-gradient contribution for Pass-1 tokens uses their sampled log probabilities and A_i^{EC}; KL regularization remains active. Thus suppressing chain advantage in state CU does not suppress the KL term.

A.2.2 Training Configuration

Table A.1: Reported model, verifier, and interface configuration. Model repository identifiers do not specify immutable checkpoint revisions.

Setting	Value
Policy	`Qwen/Qwen3-VL-8B-Instruct`
Trainable parameters	Vision, projector, and language model
Reference policy	Frozen policy initialization
Visual verifier	`alibaba-pai/VideoCLIP-XL`, BF16, 8 frames, CC-BY-NC-SA-4.0
Text verifier	`Qwen/Qwen2.5-14B-Instruct`, BF16, constrained label scoring
FutureBench frames	32
VEPBench frames	16
Maximum frame resolution	448×448
Pass-2 chain interface	Read-only (video remains available)
Pass 1 / Pass 2 limits	256 / 384 generated tokens
Maximum context	8192 tokens
Event-chain length	2–5 events
Training organization	Separate FutureBench, VEPBench, and AVEP runs

A.2.3 Budget and Interface Controls

The controlled Qwen3-VL-8B runs share the reported backbone, benchmark data, frame allowance, nominal token allowance, and update count. The answer-only GRPO baseline does not expose the same chain interface as the full method, so this comparison changes both supervision and interface. Equal nominal budgets also do not establish equal realized token counts, verifier cost, wall-clock time, or GPU-hours. The "w/o Two-Pass" and "w/o PACR" rows in the main paper test the reported component variants; they do not resolve all of these factors independently.

No results are reported for two-pass answer-only GRPO, token-matched free-form CoT, an event chain with answer reward only, or an event chain without timestamps. Consequently, the available comparisons do not isolate every benefit of the two-pass interface from its reward design and computation.

Table A.2: Reported reward, optimization, and hardware configuration. Hardware allocation alone does not determine training cost.

Setting	Value
Semantic weights	$(0.50, 0.25, 0.25)$
Structure parameters	$\lambda_{\text{fmt}} = 0.7,\ \kappa = 0.5$
Initial structural weight η_0	0.20
Minimum structural weight $\eta_{\min}$	0.05
Annealing horizon	$T_{\text{anneal}} = 0.2T$
Alignment quantiles	Per-benchmark contrastive-score $Q_{0.05}, Q_{0.95}$
Sufficiency-gain scale	Per-benchmark positive $Q_{0.95}$ of Δ_ϕ
Routing thresholds	$(\tau_F, \tau_P) = (0.60, 0.50)$
PACR coefficients	$(\rho_f, \rho_m, \rho_p, \zeta) = (.2, .1, .2, .2)$
Prompt batch / group size	$B = 8,\ G = 8$
Updates / trajectories	300 / 19,200 per benchmark
Rollout sampling	$T = 1.0,\ p = 0.95$, no top-k
Micro-batch / accumulation	1 trajectory / 8 steps
Evaluation frequency	Every 25 updates
GRPO clipping / KL	$\epsilon_{\text{clip}} = 0.2,\ \lambda_{\text{KL}} = 0.04$
LLM/projector LR	5×10^{-7}
Vision LR	1×10^{-7}
Optimizer	AdamW, $(\beta_1, \beta_2) = (.9, .95)$, $\epsilon = 10^{-8}$
Decay / clipping	Weight decay 0.01; gradient clipping 1.0
Schedule	Cosine decay, 5% warm-up
Distributed training	DeepSpeed ZeRO-3
Training hardware	8×A100 80GB

A.3 Prompt Templates and Evaluation Protocols

The following blocks reproduce the documented shared system prompt, the two policy user prompts, and the text-only predictive sufficiency verifier prompts. Braced fields are instantiated per example; the block headings identify roles and are not prompt text. Passes 1 and 2 share the system prompt. The same two policy prompts are used for training rollouts and inference. Training additionally parses the tagged outputs to compute rewards; inference removes the reward models and group sampling.

Shared system prompt (Passes 1 and 2).

```
You are a video understanding and future-event prediction
model.
Ground every statement in the observed video.  Do not invent
actions, objects, intentions, or future events that are not
supported by visible evidence.  Follow the requested output
format exactly and output no additional text outside the
specified tags.
```

Pass 1 user prompt: option-blind historical event perception.

```
Construct a chronological event chain from the observed
video.
The video duration is {VIDEO_DURATION} seconds.
Requirements:
1.  Describe only events that are visibly observed in the
video.
2.  Generate between 2 and 5 non-redundant events.
3.  Each event must contain a start time, an end time, and
one concise description.
4.  Use timestamps in seconds, with one decimal place,
within [0.0, {VIDEO_DURATION}].
5.  Arrange events chronologically.
```

```
6. Focus on meaningful state changes, object interactions,
body movements, and ongoing action precursors.
7. Describe a preparatory movement only when the movement
is already visibly occurring.
8. Do not predict what happens after the observed video.
9. Do not describe intentions, goals, or unobserved
consequences.
10. Do not repeat the same event using different wording.
Use exactly the following format:
<event>
Time: {start_time}-{end_time}
Description: {observed event}
</event>
<event>
Time: {start_time}-{end_time}
Description: {observed event}
</event>
<end_event_chain>
Output only the event chain.
```

Pass 2 user prompt: future-event prediction.

```
Predict the most likely event that occurs next after the
observed video.
You are given:
1. the observed video;
2. a historical event chain generated from the video;
3. several candidate future events.
The historical event chain is an intermediate summary and
may be incomplete or inaccurate. Verify its claims against
the observed video before using it.
Treat all candidate events as alternative future
possibilities. Do not treat any candidate event as
something that has already occurred.
Task question:
{QUESTION}
Historical event chain:
{EVENT_CHAIN}
Candidate future events:
{OPTION_LIST}
Instructions:
1. Identify the most recent observed state changes and
action precursors.
2. Use the original video as the primary evidence.
3. Explain which observed evidence supports the selected
future.
4. Briefly distinguish the selected option from the
strongest competing option.
5. Do not invent events that are not visible in the
observed video.
6. Select exactly one candidate label.
Use exactly the following format:
<reasoning>
{concise evidence-based reasoning}
</reasoning>
<answer>{OPTION_LABEL}</answer>
Output no additional text.
```

Text-only predictive sufficiency verifier: system prompt.

```
You are a strict evidence-sufficiency verifier.
Determine whether the provided historical event chain alone
contains enough explicit information to distinguish one
candidate future event from all other candidates.
Do not evaluate whether the event chain is visually correct.
Visual faithfulness is assessed separately.
Do not fill in missing evidence using general world
knowledge, common event scripts, or assumptions about what
usually happens.  If the chain is generic, ambiguous,
missing the decisive precursor, or supports multiple
candidates, output INSUFFICIENT.
Return exactly one allowed label and no explanation.
```

Text-only predictive sufficiency verifier: user prompt.

```
Task question:
{QUESTION}
Historical event chain:
{EVENT_CHAIN}
Candidate future events:
{OPTION_LIST}
Select the uniquely supported candidate using only the
historical event chain.
Allowed outputs:
{OPTION_LABELS}, INSUFFICIENT
```

For the options-only predictive control, the chain field is replaced by <none> and the task question Q, every other prompt token, and the label-scoring rule are unchanged. During training, Pass 2 and the sufficiency verifier receive independently sampled candidate permutations. Their selected display labels are mapped back to semantic candidate identities before answer or sufficiency rewards are computed.

Closed-set benchmark evaluation.

```
[VIDEO]
{observed_video}
[CANDIDATE FUTURES]
{labelled_options}
[INSTRUCTION]
Choose the candidate most likely to happen next in the
video.  Respond with only its option label.  Do not
output an explanation.
```

This prompt is used for vanilla commercial and open-source baselines; the frame budget matches the corresponding benchmark. Display order is independently randomized per instance, and the parsed display label is mapped back to the canonical semantic type before scoring. The displayed baseline template does not specify a separate task-question field; the actual serialization of Q for baseline evaluation still requires documentation.

A.4 Additional Results and Behavioral Analyses

Main-paper Tables 1 and 2 contain the complete recorded FutureBench and VEPBench comparison inventories. Model citations identify the systems rather than the provenance of every measurement; the controlled SFT/GRPO rows are evaluations in this study. All numbers remain reported estimates subject to the reconciliation in Section A.6.

A.4.1 FULL AVEP RESULTS

Tables A.3 and A.4 provide the complete reported baseline inventory for the AVEP comparison in the main paper. AVEP is trained in a separate benchmark-specific run. For each target, Pass 2 scores the finite verb, noun, or action vocabulary by constrained decoding; the answer reward is exact benchmark label match. This is a separately trained evaluation, not zero-shot transfer from the other benchmarks. Values marked "n/a" are unavailable; they are neither zero scores nor pending table entries.

Table A.3: Full AVEP verb and noun results (Su et al., 2025). T/V denotes test/validation and P/R/F1 denote precision, recall, and F1. Higher is better; bold and underlined entries are the best and second-best values. "n/a" denotes unavailable results.

Model / Method	Verb T/V	Noun Test			Noun Val		
		P	R	F1	P	R	F1
Open-Source Models							
LLaVA-Video-7B (Zhang et al., 2025b)	5.67 / 6.27	44.50	41.95	43.19	44.96	41.64	43.24
Kimi-VL-A3B (Kimi Team, 2025)	5.12 / 5.30	31.79	36.43	33.95	31.46	37.04	34.02
InternVL3-8B (Zhu et al., 2025)	5.80 / 7.00	48.77	45.77	47.22	43.64	37.94	40.59
GLM-4.1V-9B (GLM-V Team, 2025)	9.10 / 10.62	33.80	33.66	33.73	34.94	35.42	35.18
MiMo-VL-7B (Xiaomi LLM-Core Team, 2025)	10.23 / 9.57	43.46	32.07	36.91	39.54	35.10	37.19
Qwen2.5-VL-7B (Bai et al., 2025b)	5.73 / 6.77	44.83	56.54	50.01	44.26	59.33	50.70
Qwen2.5-VL-72B (Bai et al., 2025b)	8.00 / 7.95	48.94	41.38	44.84	<u>49.24</u>	41.57	45.08
Visual Reasoning							
VideoChat-R1 (Li et al., 2025b)	8.31 / n/a	n/a	n/a	42.22	n/a	n/a	n/a
Video-R1 (7B) (Feng et al., 2025)	9.47 / n/a	n/a	n/a	47.04	n/a	n/a	n/a
Qwen3-VL-8B-GRPO	10.00 / 11.00	48.00	91.00	63.00	47.80	92.00	63.80
CoE-GRPO-7B (Su et al., 2026)	<u>12.24</u> / **18.75**	<u>49.88</u>	<u>93.95</u>	<u>65.16</u>	48.68	<u>93.54</u>	<u>64.03</u>
PGE-RL (8B) (ours)	**15.00** / <u>16.00</u>	**52.00**	**95.00**	**68.00**	**51.00**	**96.00**	**69.00**

Table A.4: Full AVEP action results. T/V denotes test/validation. Higher is better; "n/a" denotes unavailable values.

Model / Method	Action P (T/V)	Action R (T/V)	Action F1 (T/V)
LLaVA-Video-7B (Zhang et al., 2025b)	3.13 / 3.57	2.95 / 3.30	3.04 / 3.43
Kimi-VL-A3B (Kimi Team, 2025)	2.40 / 2.83	3.27 / 3.16	2.77 / 2.99
InternVL3-8B (Zhu et al., 2025)	3.37 / 3.60	3.23 / 3.07	3.30 / 3.31
GLM-4.1V-9B (GLM-V Team, 2025)	3.01 / 3.27	3.02 / 3.34	3.01 / 3.30
MiMo-VL-7B (Xiaomi LLM-Core Team, 2025)	<u>6.96</u> / 5.54	5.08 / 3.86	5.87 / 4.55
Qwen2.5-VL-7B (Bai et al., 2025b)	2.21 / 3.76	3.21 / 4.64	2.62 / 4.15
Qwen2.5-VL-72B (Bai et al., 2025b)	5.32 / 5.38	4.25 / 4.41	4.72 / 4.85
VideoChat-R1 (Li et al., 2025b)	n/a / n/a	n/a / n/a	3.95 / n/a
Video-R1 (7B) (Feng et al., 2025)	n/a / n/a	n/a / n/a	4.32 / n/a
Qwen3-VL-8B-GRPO	5.20 / 6.00	9.50 / 10.00	6.70 / 7.50
CoE-GRPO-7B (Su et al., 2026)	6.62 / <u>7.70</u>	<u>11.09</u> / **13.80**	<u>8.29</u> / <u>9.88</u>
PGE-RL (8B) (ours)	**7.50 / 8.50**	**12.00** / <u>13.00</u>	**9.00 / 10.00**

A.4.2 COMPONENT DIAGNOSTICS AND EVALUATION POPULATIONS

Figure A.6 visualizes the diagnostic trade-offs in the main-paper component ablation.

Metric definitions and evaluation populations. The reported GSR of 70.1% and the Flip measurements cover FutureBench and VEPBench. Per-benchmark diagnostics, sample counts, and the rule combining the two benchmarks are not specified. GSR must be compared with answer accuracy on the same population and under the same weighting; the summary across benchmarks cannot be directly compared with VEPBench SR@A alone. The exact Align./Suff. populations also require specification.

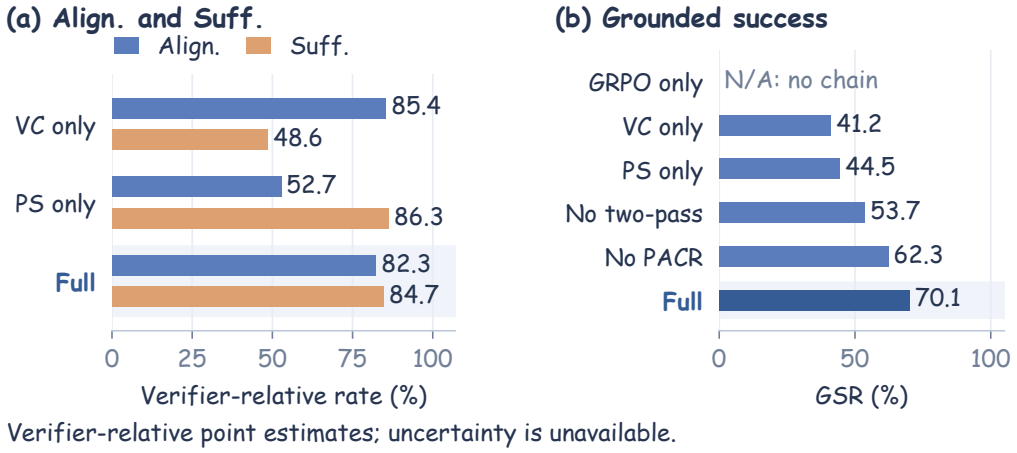


Figure A.6: Visual summary of main-paper Table 4. VC-only and PS-only training favor different verifier-relative diagnostics; the combined method has the highest reported GSR. GSR covers FutureBench and VEPBench, with unreported aggregation weights; the Align./Suff. populations remain unspecified. These point estimates do not establish statistical significance.

A.4.3 Chain-Input Intervention and Dependence Tests

Pass 2 receives the video, task question, candidates, and the read-only chain. We intervene on one input after Pass 1 while keeping the trained checkpoint and decoding fixed. Swapped chains come from a different example with similar duration, chain length, and coarse scene type; temporal corruption permutes descriptions across legal intervals; generic controls replace descriptions with scene-level text. These interventions test behavioral dependence on the chain. They do not by themselves identify a causal mechanism because corrupted inputs can introduce distribution shift.

Table A.5: Available chain-input interventions. All rows use the same trained checkpoint and decoding settings. Δ is the drop from Full and Flip is the percentage of predictions that change. The measurements support a behavioral-dependence interpretation; they are not a causal identification. Flip was evaluated on FutureBench and VEPBench; per-benchmark values, sample counts, and aggregation weights are unspecified, so the displayed values cannot be read as benchmark-specific rates.

Pass-2 condition	**FB AVG**	**VEP SR@A**	**Flip (%)**	**Δ FB**	**Δ A**	**Video**
Full $V + \mathcal{E} + \mathcal{O}$	86.10	68.95	0	0	0	yes
No chain $V + \mathcal{O}$	73.50	63.80	38	12.60	5.15	yes
Swapped chain	61.20	50.50	72	24.90	18.45	yes
Temporal corruption	65.80	54.60	63	20.30	14.35	yes
Generic chain	68.40	57.20	52	17.70	11.75	yes
Chain only $\mathcal{E} + \mathcal{O}$	70.20	61.80	48	15.90	7.15	no
Options only $\mathcal{O}$	58.20	43.70	~ 74	27.90	25.25	no

The task question Q is held fixed and omitted from the table notation. The full condition retains the video in Pass 2, whereas chain-only and options-only controls remove it.

Every reported perturbation lowers the answer metric relative to Full, which is consistent with behavioral use of the learned history. Flip measures sensitivity rather than quality. The reported Flip percentages do not establish whether a changed prediction became correct. Evaluation covers FutureBench and VEPBench, but the sample counts and combination rule are not specified. We do not label these differences causal effects.

Figure A.7 shows the reported score drops and prediction flips for the same interventions.

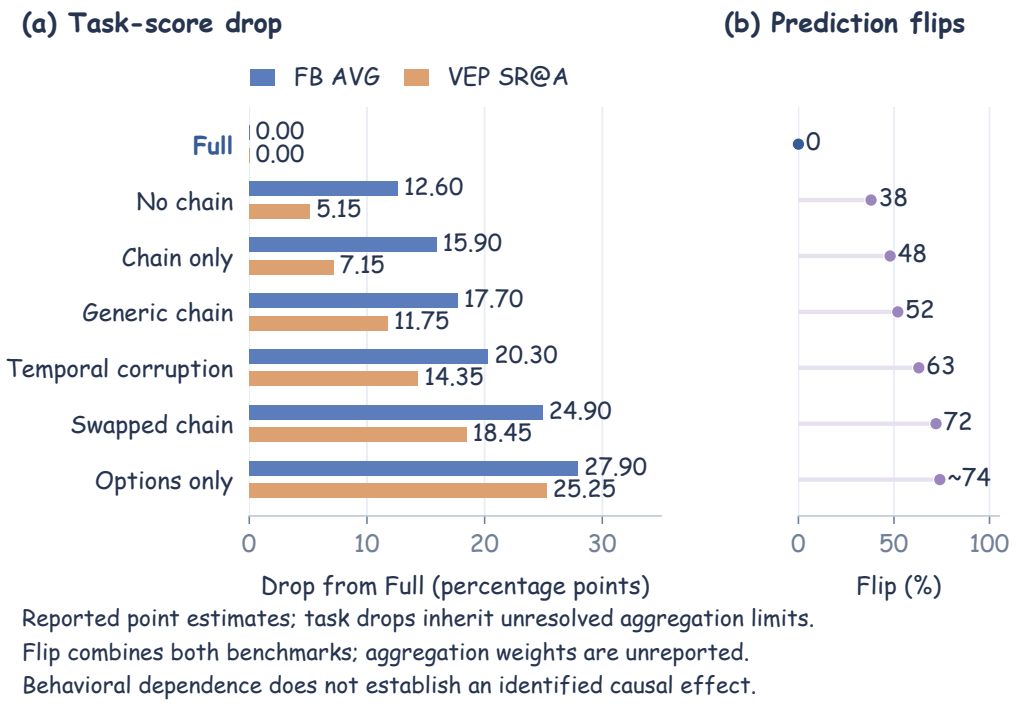


Figure A.7: Visual summary of Table A.5. Left: drops from Full on FutureBench AVG and VEP-Bench SR@A. Right: prediction flips across both benchmarks (weights unreported; options-only is approximate). Flip measures sensitivity, not whether a changed answer is correct. The interventions support behavioral dependence, subject to distribution shift and the metric-reconciliation limits in Section A.6.

A.5 Human Assessment and Verifier Reliability

Human assessment. A human audit evaluated 500 samples with three independent annotators. Agreement between the verifier-based decisions and majority human votes exceeded 90% for each criterion: VC $>$ 90%, PS $>$ 90%, and GSR $>$ 90%. This supports agreement with human judgments on the audited sample; it does not make the proxies factual guarantees. The sampling procedure, benchmark allocation, raw annotation labels, classwise results, and inter-rater kappa are not included, so neither finer-grained agreement rates nor uncertainty estimates are reported.

Verifier validity and unreported analyses. Predictive grounding is the ideal conjunction of visual consistency and predictive sufficiency; the reported Align., Suff., and GSR measurements use verifier-relative proxies. A text verifier can find a chain discriminative without establishing that the described events occurred. The reported evaluation does not include evaluator-transfer results, verifier-backbone ablations, or routing-threshold/coefficient sensitivity results. Robustness to a different verifier or routing configuration is therefore untested. The diagnostic names for the five PACR states describe reward routing, not independently verified causes of errors. These limitations also constrain interpretations of the chain interventions in Table A.5.

A.6 Evidence Boundaries and Reproducibility

Numerical reconciliation. The official FutureBench split (Wang et al., 2025a) has 1,056 examples: 173/193/201/489 for 1-Hop/2-Hop/3-Hop/interpolation. Its example-weighted AVG uses these counts. Applying them to the displayed PGE-RL category scores gives approximately 85.56, rather than the reported 86.10; several other rows also fail this reconstruction beyond rounding tolerance. The evaluation manifests and raw aggregation outputs are needed to establish whether a differ-

ent split or protocol was used. VEPBench is described as containing 3,000 examples, but 68.95% cannot be the two-decimal rounding of a single integer success count out of 3,000; the actual denominator and averaging rule need clarification. For AVEP, the reported noun test/validation F1 scores 68.00/69.00 exceed the harmonic means 67.21/66.61 of the corresponding displayed precision and recall. Averaging per-item F1 with the same nonnegative weights used for precision and recall cannot by itself explain these differences. The exact evaluator, masks, and aggregation rules are needed; no alternate rule is assumed. Split labels and baseline provenance also require reconciliation: the AVEP source reports LLaVA-Video verb scores of 5.67/6.27 for validation/test, whereas Table A.3 lists these values as test/validation. The CoE-GRPO entries follow the published comparison (Su et al., 2026); agreement with those reported values does not resolve the split-label or aggregation questions above. These remaining discrepancies require reconciliation with the original evaluation outputs. The reported values are retained without substituting diagnostic recalculations for experiment results.

Reproduction materials. Tables A.1 and A.2 provide the reported settings, but they are not a complete executable reproduction specification. The manuscript package does not include raw experiment logs, immutable policy/verifier checkpoint revisions, random seeds, the number of independent runs, or GPU-hours. AVEP's frame budget and the exact target-vocabulary/decoding configuration are also unspecified. Consequently, no confidence intervals, seed variability, statistical significance, or measured compute-equivalence claim is made here.

Implementation details requiring specification. The documented prompt blocks and the schematic rollout do not substitute for the actual serialized training traces and parser implementation. Reproduction additionally requires the exact overlap rule, construction and size of the negative-interval set, behavior when that set is empty, handling of coincident calibration quantiles or absent positive predictive gains, and masking/normalization when a generated stage is empty. The value of the advantage-normalization stabilizer ϵ_{norm} is unspecified; it is distinct from the clipping parameter $\epsilon_{\mathrm{clip}} = 0.2$. The runtime definition of R_{align} for an empty chain is also needed: specifying $F_i = 0$ does not define that mean reward. Their runtime handling remains unspecified. The training-time verifiers, reward computation, and PACR routing are removed at inference; the two policy passes and immutable chain interface remain.