

From Plausible to Grounded: Reinforcing Structured Consistency in Video MLLMs

Yihao Quan¹ Zeru Shi¹ Jinman Zhao² Ruixiang Tang^{†1}

¹Rutgers University ²University of Toronto

yq207@rutgers.edu

Abstract

Multimodal large language models (MLLMs) have achieved remarkable progress in video understanding. However, seemingly plausible outputs often suffer from poor visual and temporal grounding: a model may fabricate object existence, assign incorrect attributes, or collapse repeated events while still producing a globally reasonable caption or answer. We study this failure mode through a compositional consistency audit that decomposes a caption into supporting factual and temporal claims, investigating whether a correct high-level prediction is actually backed by valid lower-level evidence. Our top-down audit reveals that even correct root relational claims often lack reliable attribute and existence support. This indicates that standard sentence-level supervision is a weak proxy for faithful video understanding. Furthermore, when turning to reinforcement learning (RL) for better alignment, standard sentence-level rewards often prove too coarse to accurately localize specific grounding failures. To address this, we replace generic sentence-level rewards with a structured reward built from factual and temporal units. Our training objective integrates three complementary components: (1) an instance-aware scene-graph reward for factual objects, attributes, and relations; (2) a temporal reward for event ordering and repetition; and (3) a video-grounded VQA reward for hierarchical self-verification. Across temporal, general video understanding, and hallucination-oriented benchmarks, this objective yields consistent gains on open-source backbones. These results suggest that structured reward shaping is a practical route to more faithful video understanding.

1 Introduction

Multimodal large language models (MLLMs) (Liu et al., 2023; Wang et al., 2024a; Achiam et al., 2023; Team et al., 2023) have recently made strong progress in video understanding, serving as the foundation for tasks like video captioning (You et al., 2016; Xu et al., 2016), question answering (Antol et al., 2015; Yu et al., 2019), and agent-style interaction. Particularly in open-ended generation, the resulting text acts not just as a final prediction, but as an explicit reflection of the model’s underlying understanding. Consequently, factual and temporal faithfulness is paramount. While current models can produce globally plausible outputs, they frequently fail to ground these narratives in reliable visual evidence. A minor error in object identity, attribute assignment, or event ordering can invalidate an otherwise coherent caption. This exposes a critical vulnerability: the capacity of MLLMs to generate plausible text has outpaced their ability to verify the atomic facts that support it.

Top-down audit of caption consistency. To systematically investigate this failure mode, we shift the evaluation focus from surface sentence plausibility to grounded internal consistency. We conduct a top-down compositional consistency audit, decomposing a caption into a hierarchy of supporting factual and temporal claims. Starting from a root relational claim, we then check whether the prerequisite lower-level evidence is also correct. Let S

[†]Corresponding Author

denote the set of audit samples, let $S_r = \{i \in \mathcal{S} : \text{Root}_i = \text{GT}_i\}$, and let $S_{ra} = \{i \in \mathcal{S} : \text{Root}_i = \text{GT}_i \wedge \text{Attr}_i = \text{GT}_i\}$. Further let $c_i^r = \mathbf{1}[\text{Root}_i = \text{GT}_i]$, $c_i^a = \mathbf{1}[\text{Attr}_i = \text{GT}_i]$, and $c_i^e = \mathbf{1}[\text{Exist}_i = \text{GT}_i]$. We report three empirical accuracies: Root Relation Accuracy (RRA), Attribute Conditional Accuracy (ACA), and Existence Conditional Accuracy (ECA):

$$\text{RRA} = \mathbb{E}_{i \in \mathcal{S}}[c_i^r], \quad \text{ACA} = \mathbb{E}_{i \in S_r}[c_i^a], \quad \text{ECA} = \mathbb{E}_{i \in S_{ra}}[c_i^e].$$

These three metrics progressively test whether a caption remains correct as we descend the evidence hierarchy. RRA evaluates the top-level relational claim over the full sample set. ACA then asks, among samples with a correct root relation, whether the supporting attribute is also correct. ECA further asks, among samples with both a correct relation and a correct attribute, whether the referenced object existence is correct. Together, this cascade tests whether an apparently correct high-level statement is actually supported by the lower-level evidence beneath it, as illustrated in Figure 1.

Video models are not consistent captioners. Our top-down audit reveals a striking vulnerability in state-of-the-art models: a severe deficit in fine-grained grounding even when high-level predictions are correct. Logically, if a model genuinely comprehends a scene well enough to correctly predict a complex root relational claim, its accuracy on the prerequisite lower-level evidence should approach 100%. However, our conditional analysis on Qwen3-VL-8B exposes a substantial consistency gap. Even under the favorable condition where the root claim is already correct, the supporting attribute accuracy only reaches approximately 60%, and the object-existence accuracy still falls short at around 80%. Thus, a correct high-level compositional statement does not guarantee that the underlying fine-grained evidence is actively grounded. The model frequently bypasses foundational factual evidence to guess the global context, rendering the reasoning process right for the wrong reasons.

Structured consistency via reinforcement learning. Addressing this deficit requires training signals that better capture whether generated captions are grounded in the video at the level of factual and temporal consistency. We therefore revisit reinforcement learning (RL) through structured reward design. Specifically, we parse generated and reference captions into structured factual and temporal units, and use them to build a sequence-level training objective with three complementary reward branches: (1) a scene-graph reward over objects, attributes, and relations; (2) an instance-aware temporal reward for repeated-event alignment and order consistency; and (3) a video-grounded VQA reward for hierarchical verification. This design better aligns the reward with grounded consistency, while retaining standard sequence-level policy optimization.

Extensive experiments demonstrate that this objective improves caption faithfulness while remaining highly competitive on standard video understanding benchmarks. Across Qwen3-

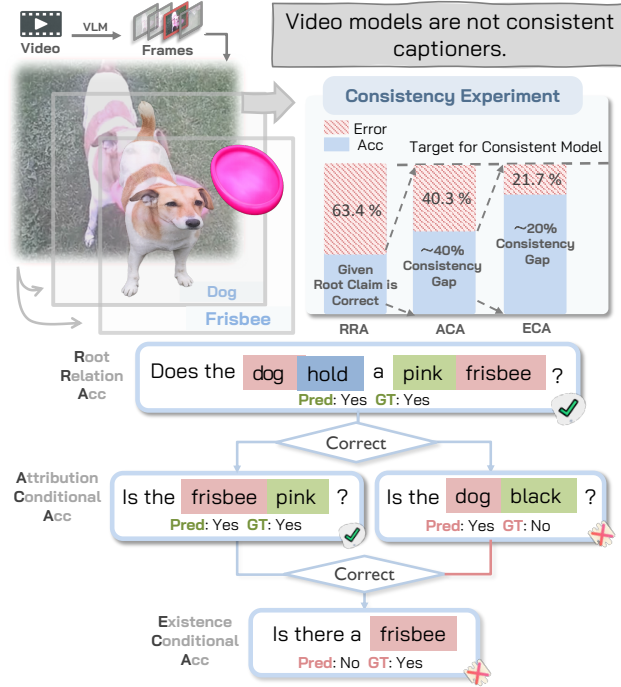


Figure 1: Motivating observation. We evaluate a caption with a top-down compositional consistency audit. Starting from a root claim, we ask whether the supporting evidence beneath it is also correct. The audit set contains 200 human-created audit samples derived from AGQA-Decomp, with the root, attribute, and existence labels manually refined and verified.

VL backbones, the strongest and most consistent gains are observed on hallucination-sensitive evaluations. This pattern matches the intended role of our method: structured consistency supervision acts as a powerful mechanism for reducing unsupported content, with downstream performance gains naturally following from this enhanced faithfulness.

In summary, our main contributions are as follows:

- We adapt compositional consistency auditing to video captioning, revealing a critical flaw in current VideoLLMs: top-down analysis shows they frequently generate plausible high-level claims that lack valid attribute and existence grounding, functioning as inconsistent visual captioners.
- Motivated by this gap, we design structured rewards for sequence-level RL in video captioning. Our objective integrates (1) an instance-aware scene-graph reward, (2) an temporal reward, and (3) a video-grounded VQA reward to explicitly enforce factual and temporal consistency.
- We demonstrate that optimizing structured consistency yields robust improvements across multiple open-source backbones. Notably, our approach delivers the most significant gains on hallucination-sensitive evaluations, validating it as a highly practical route to faithful video understanding.

2 Related Work

Compositional reasoning in video understanding. A central idea behind our method is that video understanding should be judged compositionally, not only by a final sentence or answer. This line is most explicit in VideoQA. AGQA introduced compositional spatio-temporal reasoning as a benchmark target, and AGQA-Decomp showed that even when a final answer is correct, its supporting sub-claims may still be inconsistent (Grunde-McLaughlin et al., 2021; Gandhi et al., 2022). More recent frameworks attempt to expose intermediate structure directly: Video-of-Thought organizes reasoning from perception to cognition, while Align-and-Aggregate parses a question into a compositional graph and resolves the final answer through intermediate sub-questions (Fei et al., 2024; Liao et al., 2024). In parallel, grounded and causal VideoQA methods try to isolate the evidence relevant to each question by alignment or causal refinement (Li et al., 2022; Wei et al., 2023; Chen et al., 2025). Our work is closest in spirit to this literature, but the setting is different: instead of decomposing a *question* at inference time, we decompose a generated *caption* during training and use its scene-graph atoms and ordered events as reward-bearing supervision targets.

3 Method

3.1 From captions to anchored structured units

Given a video-instruction pair $[V, x]$, the policy samples a caption $y \sim \pi_\theta(\cdot \mid [V, x])$, where y^* denotes the reference caption. Before computing rewards, we first segment each caption into sentences and clauses using spaCy, and normalize noun and verb heads by lemmatization (Honnibal et al., 2020). We then apply the Factual scene-graph parser (Li et al., 2023) to map both captions into a shared structured space. Formally, we define a parsing and anchoring function f_{parse} that extracts a structured unit set $U(c) = \{U_{\text{obj}}(c), U_{\text{attr}}(c), U_{\text{rel}}(c), U_{\text{event}}(c)\}$ for any caption $c \in \{y, y^*\}$. Since the parser itself is not instance-aware, we further introduce a deterministic anchoring layer on top of its outputs. Concretely, repeated mentions are assigned local instance ids, such as cup_1 and cup_2 , and each attribute, relation, and event mention inherits the ids of the object instances on which it depends. After this step, each caption is represented as anchored objects, anchored attributes, anchored relations, and anchored event mentions, together with their participant anchors and clause positions.

These anchors provide the basis for instance awareness: two structured units $u \in U(y)$ and $u^* \in U(y^*)$ are considered matchable only when both their semantics and their anchors are

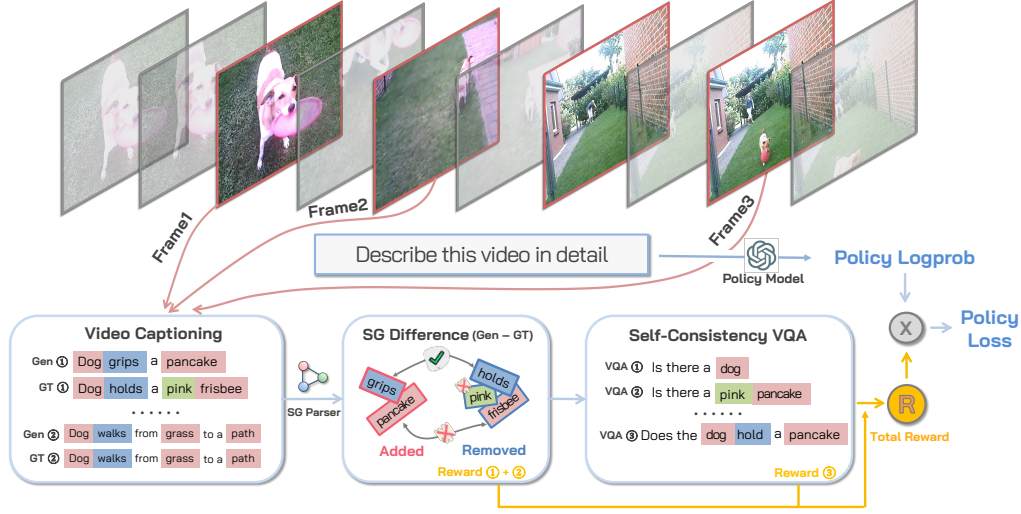


Figure 2: Overview of our method. We compare the sampled caption and the reference caption in a shared structured space rather than at the whole-sentence level. After lightweight preprocessing, both captions are parsed into scene-graph elements, scored by revision-based factual, temporal, and VQA verification rewards, and then used to update the policy with REINFORCE plus KL regularization.

compatible. This design allows the reward to distinguish repeated entities and repeated events without introducing any additional learned module. Full implementation details are provided in the Appendix.

3.2 Structured consistency rewards

Our reward keeps the same three-branch design, but the three branches now play cleaner and more distinct roles. The scene-graph branch scores structured revisions between the sampled caption and the reference caption. The temporal branch and the VQA branch both use the same video-conditioned verification interface, defined as $\hat{a}(q) = \arg \max_{a \in \{\text{yes}, \text{no}\}} \pi(\cdot \mid [V, q])$, but they supervise different root claims: the temporal branch targets event order and repetition, while the VQA branch targets factual support chains over objects, attributes, relations, and event participants.

Revision-based scene-graph reward. We first align generated and reference object instances by finding a maximum-weight one-to-one bipartite matching $\mathcal{M}_{\text{obj}}$ between $U_{\text{obj}}(y)$ and $U_{\text{obj}}(y^*)$, where the edge weight is the semantic similarity between canonicalized object phrases. This object map $\mathcal{M}_{\text{obj}}$ defines which higher-order units are allowed to match. An attribute can match only when it is attached to the same aligned object, a relation can match only when both endpoints are aligned, and an event mention can match only when its participant anchors are compatible under the same object map. For each unit type $\tau \in \{\text{obj}, \text{attr}, \text{rel}\}$, we remove exact overlaps and then solve a second partial one-to-one matching on the residual units. Let m_τ denote the resulting matched mass, including exact overlaps and the similarity of valid residual substitutions. We score each unit type by

$$q_\tau = \frac{2m_\tau}{|U_\tau(y)| + |U_\tau(y^*)|}, \quad q_{\text{sg}} = \sum_{\tau \in \{\text{obj}, \text{attr}, \text{rel}\}} \alpha_\tau q_\tau,$$

where $\sum_\tau \alpha_\tau = 1$. If a unit type is absent on both sides, we set its score to 1. This revision-based design blocks the main reward-hacking case: once one reference unit has been claimed by one generated unit, extra near-synonymous additions can no longer collect reward and only enlarge the denominator.

Temporal reward as temporal root-claim verification. We keep temporal as an explicit reward branch, but express it with the same yes/no verification form as VQA. The key difference from the factual VQA branch is that the root claim is now temporal, not relational. We first match generated and reference event mentions one-to-one under predicate similarity after checking participant-anchor compatibility under $\mathcal{M}_{\text{obj}}$. We formulate a temporal question set $Q_{\text{temp}} = Q_{\text{temp}}^+ \cup Q_{\text{temp}}^-$. Temporal questions are instantiated only when the reference caption explicitly expresses temporal structure. Positive temporal questions Q_{temp}^+ come from reference-supported event occurrences and ordered event pairs. Negative temporal questions Q_{temp}^- are restricted to two clean counterfactuals. The first is *order inversion*: if the reference states that one matched event happens before another, we also ask the reversed order question. The second is *instance collapse or swap*: if the sampled caption merges repeated events into the same actor, or assigns an event to the wrong actor or object anchor, we convert that conflicting temporal claim into a negative question, but only within the same matched event slots. This keeps the branch focused on genuine temporal errors rather than open-world omissions.

For each temporal question $q \in Q_{\text{temp}}$, we query the same video-conditioned policy with $[V, q]$, read a detached one-step answer $\hat{a}(q) \in \{\text{yes}, \text{no}\}$, and compare it with the known label $a^*(q)$. The temporal score is the average binary accuracy:

$$q_{\text{temp}} = \frac{1}{|Q_{\text{temp}}|} \sum_{q \in Q_{\text{temp}}} \mathbf{1}[\hat{a}(q) = a^*(q)].$$

Hierarchical factual VQA reward. The VQA branch uses the same binary verification interface, but its goal is different: it checks whether high-level factual claims are supported by the lower-level evidence underneath them. Starting from the reference parse, we expand each supported root claim into a short support chain over object existence, anchor-specific attributes, anchored relations, and event participants. These form the positive question set Q_{vqa}^+ . The negative question set Q_{vqa}^- does not consist of arbitrary facts absent from the reference caption. Instead, we only use anchor-compatible conflicts at the same structural slot, such as a wrong attribute on a matched object, a wrong relation between matched objects, or a wrong participant binding exposed by the sampled caption. This restriction is important because it avoids penalizing valid but unmentioned video content.

Letting $Q_{\text{vqa}} = Q_{\text{vqa}}^+ \cup Q_{\text{vqa}}^-$, we use the same detached yes/no answering step as above and define:

$$q_{\text{vqa}} = \frac{1}{|Q_{\text{vqa}}|} \sum_{q \in Q_{\text{vqa}}} \mathbf{1}[\hat{a}(q) = a^*(q)].$$

Detailed templates, filtering rules, and positive/negative construction rules will be shown in the appendix.

We convert all normalized branch scores into centered rewards with a shared affine form

$$r_b = \rho (q_b - \kappa), \quad b \in \{\text{sg}, \text{temp}, \text{vqa}\},$$

and combine them as

$$R = \lambda_{\text{sg}} r_{\text{sg}} + \lambda_{\text{temp}} r_{\text{temp}} + \lambda_{\text{vqa}} r_{\text{vqa}}.$$

Here κ has the shared centering constant value 0.5 and ρ controls the reward range. We keep these values fixed in the main experiments and report the search ranges and robustness ablations in the Appendix.

3.3 Policy optimization

We optimize the policy with a REINFORCE-style objective over sampled captions. Let $\bar{\ell}_\theta(y) = \frac{1}{L} \sum_{t=1}^L \log \pi_\theta(a_t | a_{<t}, [V, x])$ be the average log-probability of the sampled caption and let $\text{KL}(\pi_\theta \| \pi_{\text{ref}})$ be the sampled-token KL against a frozen reference policy. The training loss is defined over the expectation of the generated samples $y \sim \pi_\theta$:

$$\mathcal{L}(\theta) = -\mathbb{E}_{y \sim \pi_\theta} [R \cdot \bar{\ell}_\theta(y)] + \beta \mathbb{E}_{y \sim \pi_\theta} [\widehat{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})].$$

Fundamentally, our design functions as structured reward shaping for sequence-level RL. By anchoring rewards to key compositional elements, it provides finer-grained supervisory signals than standard sentence-level RL, while remaining significantly more computationally efficient and stable than dense token-level reward assignment. All parser, preprocessing, matching, and reward hyperparameters are fixed across the main experiments and detailed in Appendix 2.1 and Appendix 2.2.

4 Experiments

4.1 Experimental setup

Implementation details. Unless otherwise specified, we use 8 sampled frames at inference time and keep the evaluation protocol identical across methods. For open-source backbones, we initialize from the corresponding instruction-tuned checkpoint and perform parameter-efficient post-training with LoRA adapters (Hu et al., 2022). In the default 8B recipe, we first run SFT on the 1,891-sample TemporalBench (Cai et al., 2024) short-caption set and then apply our RL stage on the aligned training set used by our method, which provides scene-graph, temporal, and QA-style supervision derived from the reference caption. Both stages are optimized with AdamW (Loshchilov & Hutter, 2019). Appendix 2.1 summarizes the default recipe, Appendix 2.2 lists the parser and preprocessing settings used by the reward modules, Appendix 3 summarizes temporal and factual question construction, Appendix 4.1 reports the parser-backend audit that motivates our choice of FACTUAL, and Appendix 4.2 audits train-evaluation overlap for the TemporalBench-based training pools.

Baselines. We compare our method against two categories of baselines: (i) **Closed-source VLMs**: GPT-4o (Achiam et al., 2023) and Gemini 1.5 Pro (Team et al., 2023). (ii) **Open-source VLMs**: LLaVA-OV-7B (Zhang et al., 2024b), Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-4B (Bai et al., 2025a), and Qwen3-VL-8B. For open-source models, we additionally compare against recent RL-based or self-correction-based variants, including Season (Wu et al., 2025), TPO (Li et al., 2025b), RRPO (Sarkar & Etemad, 2025), DINO-HEAL (Li et al., 2025a), and ArrowRL (Xue et al., 2025). Appendix 3.1 summarizes the concrete settings used for the reproduced baselines in our comparisons.

Benchmarks. We evaluate on three benchmark groups. First, we use **temporal understanding** benchmarks, TempCompass (Liu et al., 2024) and TVBench (Cores et al., 2024), to test event order and temporal reasoning. Second, we use **general video understanding** benchmarks, Video-MME (Fu et al., 2025) and MVBench (Li et al., 2024), to measure whether gains in faithfulness preserve broad video QA ability. Third, we use **hallucination-oriented** benchmarks, VideoHalluciner (Wang et al., 2024b), and EventHallusion (Zhang et al., 2024a), to probe factual and temporal failure modes more directly.

4.2 Main results

Metric	Base	Ours	Δ
Root Relation Accuracy	36.6	47.8	+11.2
Attribute Conditional Accuracy	59.7	78.4	+18.7
Existence Conditional Accuracy	78.3	92.1	+13.8

Table 1: **Post-training top-down audit under the same protocol as Figure 1.** Our method improves the full root-to-attribute-to-existence support chain, with especially large gains on the conditional attribute and existence accuracies.

Top-down audit as a matched diagnostic. The top-down audit in Section 1 is also the diagnostic that our method is designed to improve. Table 1 applies the same root-to-attribute-to-existence protocol to compare the base model against our post-trained model. The improvement is substantial at every level of the hierarchy. The root relation accuracy

Models	Temporal			Conventional			Hallucination		
	TempComp.	TVBench	Avg.	Video-MME	MVBench	Avg.	VidHalluciner	EventHall.	Avg.
<i>Closed-source Models</i>									
GPT-4o	73.8	39.9	56.9	71.9	49.1	60.5	53.3	91.9	72.6
Gemini 1.5 Pro	67.1	47.6	57.4	75.0	60.5	67.8	37.8	80.4	59.1
<i>Open-source Models</i>									
LLaVA-OV-7B	68.3	42.8	55.6	50.6	54.4	52.5	46.4	60.2	53.3
+TCD	68.5	42.9	55.7	50.4	54.3	52.4	48.1	68.5	58.3
+DINO-HEAL	68.3	42.9	55.6	50.6	54.4	52.5	46.9	60.2	53.6
+Season	69.3	43.0	56.2	50.7	54.8	52.8	49.3	69.2	59.3
+OURS	68.8	43.0	55.9	51.2	54.6	52.9	48.5	64.2	56.4
Qwen2.5-VL-7B	72.9	46.7	59.8	49.9	61.8	55.9	53.1	63.3	58.2
+TCD	73.4	46.1	59.8	49.4	60.1	54.8	52.8	64.8	58.8
+DINO-HEAL	73.2	46.9	60.1	49.9	61.9	55.9	53.3	63.6	58.5
+ArrowRL	72.6	49.4	61.0	49.6	59.8	54.7	54.5	69.0	61.8
+Season	73.7	47.7	60.7	50.4	62.3	56.4	54.8	66.0	60.4
+OURS	73.8	48.6	61.2	50.9	62.4	56.7	54.8	65.8	60.3
Qwen3-VL-4B	71.0	41.7	56.6	53.6	62.7	58.2	51.4	68.4	59.9
+Season	70.6	43.2	56.9	53.8	64.3	59.1	54.2	67.9	61.1
+OURS	73.0	47.1	60.1	54.4	64.2	59.3	55.8	72.0	63.9
Qwen3-VL-8B	74.1	47.0	60.6	56.7	65.8	61.3	55.7	66.8	61.3
+Season	71.2	46.8	59.0	56.6	65.7	61.2	53.3	71.6	62.5
+OURS	75.1	48.0	61.3	57.4	66.7	62.1	56.8	73.4	65.1
Qwen3-VL-32B	77.2	51.3	64.3	60.1	69.5	64.8	62.0	77.0	69.5
+Season	74.1	49.5	61.8	60.9	69.6	65.3	55.1	71.6	63.4
+OURS	77.3	51.4	64.4	60.3	69.8	65.1	62.5	78.0	70.3

Table 2: Performance comparisons on benchmarks for temporal, conventional, and hallucination-oriented video understanding. **Bold** marks the best per group; highlights indicate the best benchmark results.

increases markedly, which shows that the model becomes more reliable even at the top-level compositional claim. More importantly, the conditional attribute and existence accuracies improve even more, indicating that the gain is not just better global guessing: once the high-level relation is correct, the lower-level support chain is also much more likely to remain grounded. This matched audit closes the loop between motivation and evaluation: the same hierarchy that reveals the original inconsistency gap is also where our method delivers one of its clearest gains.

General video understanding evaluation. Table 2 reports results on temporal and conventional video-understanding benchmarks. The main pattern is that our training objective generalizes across multiple open-source backbones, rather than benefiting only a single model family. On LLaVA-OV-7B, our method yields a modest but consistent improvement over the base model and remains competitive with recent state of art methods. On Qwen2.5-VL-7B, it achieves the best average performance within the group. The same trend is even clearer on Qwen3-VL backbones, it improves on all four benchmarks simultaneously. Overall, these results suggest that the proposed rewards improve temporal faithfulness without sacrificing broader video-understanding ability.

Hallucination benchmark evaluation. Table 2 also shows that the gains are more pronounced on hallucination-oriented evaluation. For Qwen2.5-VL-7B, our method improves the reported VideoHalluciner score over prior baselines and remains competitive on EventHalluciner. For both Qwen3-VL-4B and Qwen3-VL-8B, it substantially improves VideoHalluciner, and EventHalluciner over the corresponding base models. This pattern indicates that the structured reward strengthens factual grounding and temporal consistency.

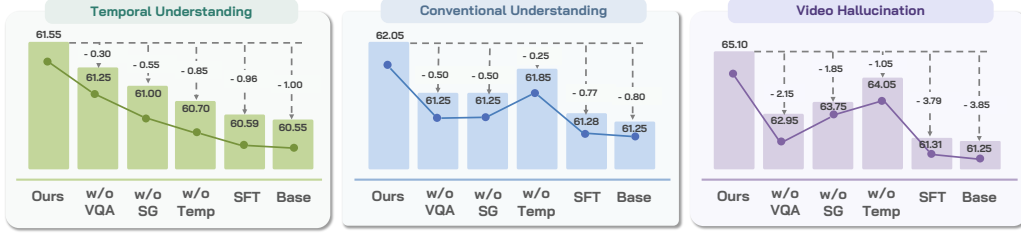


Figure 3: **Reward-component and training-stage ablations.** Results are averaged within temporal understanding, conventional video understanding, and hallucination-oriented benchmark groups.

4.3 Ablation study

We ablate the method along four axes: reward components, reward weights, training data scale, and frame sampling strategy. These ablations test complementary questions: which reward branch is necessary, whether the method is sensitive to reweighting, whether the improvement depends on large-scale training data, and whether the gains can be explained simply by changing the visual sampling budget.

Ablation on reward components. Figure 3 summarizes component and training-stage ablations by averaging benchmarks within temporal understanding, conventional video understanding, and hallucination-oriented evaluation. The full model is best in all three groups, which shows that the gain does not come from one dominant branch alone. Removing the scene-graph reward mainly hurts temporal and conventional understanding, while removing the VQA reward causes the largest drop on hallucination-oriented benchmarks. Removing the temporal reward also degrades the result, although the drop is smaller than removing the other two branches. The same figure shows that both the SFT model and the raw backbone remain clearly below the full RL system, so the final gain is not explained by supervised initialization alone. For the corresponding numeric table, see Appendix 5.3.

Setting	w_{SG}	w_{Temp}	w_{VQA}	Temporal	Conventional	Hallucination	Avg.
Ours	0.15	0.25	0.35	61.55	62.05	65.10	62.90
Rebalance to SG	0.20	0.25	0.35	61.43	62.10	65.22	62.92
Rebalance to Temp	0.15	0.30	0.35	61.70	61.96	65.00	62.89
Rebalance to VQA	0.15	0.25	0.40	61.31	62.14	65.21	62.89

Table 3: **Ablation study on reward weights for Qwen3-VL-8B.** SG refers to the scene graph reward. Results are averaged within temporal understanding, conventional video understanding, and hallucination-oriented benchmark groups.

Ablation on reward weight. Table 3 studies how sensitive the method is to reward reweighting on Qwen3-VL-8B. The main observation is that the method is robust: across all completed settings, the group-level average stays in a very narrow band so moderate perturbations of the three weights do not materially change the final result. The changes that do appear are also intuitive. Increasing the temporal weight gives the strongest temporal-group score, while increasing the VQA weight slightly favors the hallucination-oriented and conventional groups. Reweighting toward the base reward gives the highest overall average by a negligible margin, which suggests that no single branch needs to dominate the objective. Overall, the gain comes from combining all supervision signals rather than from fragile tuning of one particular weight.

Ablation on data scaling. Figure 4 shows a clear data-scaling trend: as the amount of training data increases, performance also improves across temporal understanding, conventional video QA, and hallucination-oriented evaluation. At the same time, the improvement exhibits diminishing returns. Increasing the training set from the smallest to the medium scale brings the most visible gains, while further scaling to the largest setting still improves performance but by a smaller margin. The effect is most pronounced on hallucination-

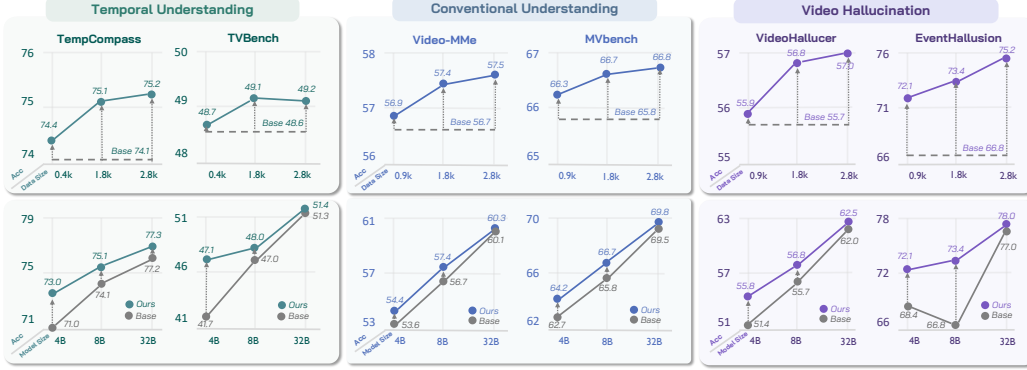


Figure 4: **Scaling behavior across model size and training data size.** We study two axes of scaling for our post-training recipe: backbone size and training data size. Results are averaged within temporal understanding, conventional video understanding, and hallucination-oriented benchmark groups.

sensitive benchmarks, with smaller yet still positive gains on temporal and conventional benchmarks. This suggests that additional training data continues to strengthen factual grounding and temporal consistency, although the marginal utility gradually decreases as the training set grows. For the complete results, see Appendix 5.2, Table 12.

Ablation on frame sampling strategy. Table 4 demonstrates that our method achieves larger performance gains when increasing the frame count from 8 to 32, indicating its ability to leverage additional visual evidence effectively. At 64 frames, however, the marginal improvement diminishes as extra frames become redundant, weakly relevant, or distracting. Overall, while a higher frame budget enhances observability, the benefits saturate when additional context adds little valuable information. Notably, the performance gap between our method and the base model remains wider at higher frame budgets, confirming that our post-training objective enables more reliable utilization of visual evidence.

Category	Benchmark	Base			Ours		
		8	32	64	8	32	64
Temporal	TempCompass	74.1	74.6	74.2	75.1 _{+1.0}	75.5 _{+0.9}	76.4 _{+2.2}
	TVBench	47.0	46.8	47.9	48.0 _{+1.0}	48.7 _{+1.9}	49.2 _{+1.3}
Conventional	Video-MME	56.7	64.2	67.0	57.4 _{+0.7}	64.6 _{+0.4}	67.3 _{+0.3}
	MVBench	65.8	68.5	69.2	66.7 _{+0.9}	69.1 _{+0.6}	69.4 _{+0.2}
Hallucination	VideoHalluciner	55.7	60.3	62.2	56.6 _{+0.9}	67.2 _{+6.9}	63.1 _{+0.9}
	EventHallusion	66.8	72.9	72.6	72.9 _{+6.1}	75.1 _{+2.2}	73.1 _{+0.5}

Table 4: **Frame-budget ablations on Qwen3-VL-8B.** Increasing sampled frames from 8 to 32 yields substantial gains across most benchmarks, while further scaling to 64 frames leads to performance saturation or slight degradation due to redundancy and noise.

5 Conclusion

In this work, we investigate an overlooked failure mode in video MLLMs: models frequently generate fluent and contextually plausible captions that lack faithful visual grounding. Our top-down compositional consistency audit reveals that even when high-level relational claims appear correct, they are often unsupported by valid underlying attribute and object existence evidence, highlighting the fundamental limitation of relying solely on standard sentence-level supervision. To address this issue, we propose a sequence-level reinforcement learning framework that enforces structured consistency by incorporating instance-aware scene-graph rewards, temporal rewards, and hierarchical VQA verification. Extensive

evaluations across temporal, general, and hallucination-sensitive benchmarks demonstrate that our approach significantly reduces object-behavior hallucinations and improves factual faithfulness without sacrificing general video reasoning performance. Overall, our findings underscore structured consistency as a key principle for building robust and trustworthy video language models, offering a promising foundation for future post-training strategies.

Acknowledgements

This work was supported in part by Advanced Micro Devices, Inc. under the AMD University Program’s AI & HPC Cluster.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.
- Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, et al. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024.
- Weixing Chen, Yang Liu, Binglin Chen, Jiandong Su, Yongsun Zheng, and Liang Lin. Cross-modal causal relation alignment for video question grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24087–24096, 2025.
- Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Lost in time: A new temporal benchmark for videollms. *arXiv preprint arXiv:2410.07752*, 2024.
- Hao Fei, Shengqiong Wu, Wenbo Ji, Hua Zhang, Meishan Zhang, Mong Li Lee, and Winston Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *International Conference on Machine Learning*, 2024.
- Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025.
- Mona Gandhi, Mustafa Omer Gul, Eva Prakash, Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Measuring compositional consistency for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5046–5055, 2022.
- Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11287–11297, 2021.

- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python, 2020. URL <https://doi.org/10.5281/zenodo.1212303>.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.
- Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13723–13733, 2025a.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.
- Rui Li, Xiaohan Wang, Yuhui Zhang, Orr Zohar, Zeyu Wang, and Serena Yeung-Levy. Temporal preference optimization for long-form video understanding. *arXiv preprint arXiv:2501.13919*, 2025b.
- Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2928–2937, 2022.
- Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. FACTUAL: A benchmark for faithful and consistent textual scene graph parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6377–6390, 2023. URL <https://aclanthology.org/2023.findings-acl.398>.
- Zhaohu Liao, Jiangtong Li, Li Niu, and Liqing Zhang. Align and aggregate: Compositional reasoning with video alignment and answer aggregation for video question-answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13395–13404, 2024.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.
- OpenAI. Openai api pricing. <https://openai.com/api/pricing/>, 2026. Accessed: 2026-03-31.
- Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 3982–3992, 2019. URL <https://aclanthology.org/D19-1410>.
- Pritam Sarkar and Ali Etemad. Self-alignment of large video language models with refined regularized preference optimization. *arXiv preprint arXiv:2504.12083*, 2025.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

- Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohalluciner: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024b.
- Yushen Wei, Yang Liu, Hong Yan, Guanbin Li, and Liang Lin. Visual causal scene refinement for video question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 377–386, 2023.
- Chang-Hsun Wu, Kai-Po Chang, Yu-Yang Sheng, Hung-Kai Chung, Kuei-Chun Wang, and Yu-Chiang Frank Wang. Season: Mitigating temporal hallucination in video large language models via self-diagnostic contrastive decoding. *arXiv preprint arXiv:2512.04643*, 2025.
- Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.
- Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. *arXiv preprint arXiv:2506.03340*, 2025.
- Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4651–4659, 2016.
- Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9127–9134, 2019.
- Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, Xingjun Ma, and Jingjing Chen. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*, 2024a.
- Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

Appendix

This appendix is organized into three parts. We first summarize the default recipe and parser configuration used in the main experiments. We then analyze two practical aspects of the training pipeline: why we use FACTUAL as the default parser and why the TemporalBench-based training pool does not create direct benchmark leakage. Finally, we collect additional captioning, scaling, and ablation results that complement the main text.

1 Declaration of LLM use

In preparing this submission, we used LLMs as writing and editing assistants for language refinement, grammar correction, formatting cleanup, and presentation polishing. They were also used to help inspect bibliography metadata and LaTeX consistency under author supervision. We did not use LLMs to generate experimental results or to replace author judgment in problem formulation, methodology, implementation, or evaluation. All substantive research decisions, experiments, and final validation were carried out and verified by the authors.

2 Training Recipe and Parser Settings

2.1 Detail Settings

Our default 8B recipe uses a simple two-stage schedule. We first apply LoRA SFT (Hu et al., 2022) on the 1,891-sample TemporalBench short-caption set and then continue with LoRA RL on the aligned training set used by our method, derived from the same caption pool. This design keeps the underlying video domain fixed across stages and changes only the supervision granularity, so the gains in the main paper are not attributable to a shift in training videos or prompt format.

Table 5 summarizes the default configuration used in the main 8B runs. In the data-scale ablation, we enlarge only the RL-stage caption pool while keeping the backbone, SFT initialization, and evaluation protocol fixed; this makes the scaling trend in Section 4 easier to interpret as a supervision-scale effect rather than a training-recipe change.

Setting	SFT	RL
Backbone	Qwen3-VL-8B-Instruct	SFT-initialized Qwen3-VL-8B
Training data	TemporalBench short-caption	Aligned SFT version of the same pool
# training samples	1,891	1,891
Fine-tuning	LoRA ($r=8, \alpha=16, \text{drop}=0.05$)	LoRA ($r=8, \alpha=16, \text{drop}=0.05$)
Optimizer	AdamW	AdamW
Learning rate	1×10^{-5}	5×10^{-6}
Schedule	cosine, warmup 5%	cosine, warmup 3%
Batch / accum.	1 / 2	1 / 4
Context length	8192	8192
Epochs / steps	1 epoch	300 steps
Rollout decoding	–	one-turn, $T=1.0, p=0.9, \text{max_new_tokens}=256$
Reward weights	–	$(w_{\text{sg}}, w_{\text{temp}}, w_{\text{vqa}}) = (0.15, 0.25, 0.35)$

Table 5: Default 8B training recipe used in the main results.

2.2 Parser and Preprocessing Settings

Our reward modules use a fixed parser stack throughout all main experiments. For scene-graph extraction, we use the public FACTUAL parser checkpoint (Li et al., 2023) ‘lizhuang144/flan-t5-base-VG-factual-sg’ and parse sentences independently before merging sentence-level outputs. This choice gives us a stable object–attribute–relation interface for the factual reward without introducing another trainable component into the RL loop.

For preprocessing, we use spaCy (Honnibal et al., 2020) with the English pipeline ‘en_core_web_sm’ to perform sentence splitting and identify noun- and verb-centered phrases

Module / Parameter	Setting
Scene-graph parser	FACTUAL parser (flan-t5-base-VG-factual-sg)
Beam size	5
Max input length	512
Max output length	128
Sentence preprocessing	spaCy (en-core-web-sm) sentence splitting

Table 6: Parser and preprocessing settings used by the reward modules in the main experiments.

and instance indexing. Repeated mentions are indexed so that downstream rewards can distinguish, for example, ‘cup.1’ from ‘cup.2’ or two repeated event mentions of the same predicate. For soft semantic comparison, we encode structured atoms with ‘sentence-transformers/all-mpnet-base-v2’ following the Sentence-BERT framework (Reimers & Gurevych, 2019). Temporal preprocessing then applies lightweight regex-based cleanup before event extraction and bucket construction. Table 6 lists the concrete parser, preprocessing, and decoding parameters used in the main runs.

3 Question construction for temporal and factual verification

The temporal and factual verification branches use the same answer interface and the same policy, but their question sources are different. Temporal questions are built only from reference event occurrences and reference event pairs that explicitly encode order or repetition. Factual questions are built from local support chains under reference objects, attributes, relations, and event participants. In both branches, the answer label is known by construction, so no external verifier is needed.

Positive questions always come from the reference parse. Negative questions are created only when the sampled caption changes a matched slot. For the factual branch, this means a wrong attribute on a matched object, a wrong relation between matched objects, or a wrong participant binding on a matched event slot. For the temporal branch, this means either a reversed supported order or a collapse/swap error where repeated events are assigned to the wrong matched actor or object anchors. We never create negatives from arbitrary extra nouns or arbitrary facts that are merely unmentioned in the reference caption.

Table 7 gives the fixed template inventory used in the revised method. In the main setting, we keep the wording fixed across all experiments and score only the first answer token over yes/no. If a hard question budget is needed, we apply an optional cap after balancing positive and negative questions.

Root type	Template	Positive source	Negative source
Object existence	Is there a/an [OBJECT]?	reference support chain	none; we do not create open-world existence negatives
Attribute	Is the [OBJECT] [ATTRIBUTE]?	reference support chain	wrong attribute on a matched object slot
Relation	Does the [SUBJECT] [RELATION] the [OBJECT]?	reference support chain	wrong relation on matched endpoints
Event occurrence	Does [PARTICIPANTS] [EVENT]?	reference event occurrence	wrong participant binding on a matched event slot
Temporal order	Did [EVENT 1] happen before [EVENT 2]?	reference ordered pair	reversed supported order; event collapse/swap conflict

Table 7: Question templates and supervision sources for the revised verification rewards. The local descriptor $d(o)$ uses the reference phrase for the matched instance and may include a local modifier or index token when it is needed to distinguish repeated instances.

3.1 Implementation details of other baselines

For reproduced non-training baselines, we use the same backbone, frame budget, and benchmark protocol as in our main runs and vary only the method-specific decoding or post-hoc scoring knobs. This keeps the comparison focused on the baseline mechanism itself rather than on unrelated prompt or sampling changes.

For TCD (Zhang et al., 2024a), we search a small grid over frame downsampling and contrastive-decoding coefficients. Concretely, we use $r \in \{2, 4\}$ and $(\alpha, \beta) \in \{(1.0, 0.1), (0.5, 0.5)\}$, which gives four configurations in total; for each benchmark, we

report the best configuration from this grid. For DINO-HEAL (Li et al., 2025a), we search over the two binary choices emphasized by the paper: whether feature normalization is enabled, and whether the DINO backbone uses registers. This again gives four configurations, and we report the best one on each benchmark.

For SEASON (Wu et al., 2025), we reproduce the method locally for Qwen3-VL and keep the main decoding coefficients fixed across benchmarks. In our runs, we use $n_{\text{frame}} = 8$, $\text{max_new_tokens} = 128$, $\alpha = 1.0$, $\beta = 0.33$, and $\text{noise std.} = 0.05$. The only backbone-specific change is the choice of self-diagnostic attention layers: we use the last four decoder layers of each model, namely 20–23 for Qwen3-VL-4B and Qwen3-VL-8B, and 36–39 for Qwen3-VL-32B. We omit training-based baselines that are evaluated directly from released checkpoints, since their comparison does not depend on a local reimplementation choice.

4 Training Pipeline Analysis

4.1 Why FACTUAL as the default parser

We additionally ran a small backend audit to decide whether the reward pipeline should use a local factual parser or an API parser. Because our pipeline parses each caption sentence by sentence before merging sentence-level outputs, we measure parser quality at the sentence level and throughput at the caption level under the same sentence-splitting pipeline. Specifically, we compare the FACTUAL checkpoint (Li et al., 2023) against GPT-5.4 mini and report the result in Table 8. For quality, we sample 48 training sentences evenly from the six source collections in the TemporalBench short-caption pool and score the resulting graphs with GPT-5.4. Since the factual reward in our final system is driven mainly by object/attribute revisions rather than relation-level graph scoring, the audit focuses on object pass, attribute pass, and an overall atom-pass criterion that asks whether a parse is usable for object/attribute supervision.

Table 8 shows that GPT-5.4 mini is the stronger parser on this small audit, especially on attributes, but FACTUAL remains usable for the factual atoms that our reward consumes. On the efficiency side, the local FACTUAL parser is $6.2\times$ faster than the API parser when measured on 100 training captions with the same sentence-by-sentence pipeline; extrapolated to the full 1,891-caption training pool, this reduces a single parsing pass from 109.1 to 17.7 minutes. Using the observed token usage together with the public GPT-5.4 mini pricing in March 2026 (OpenAI, 2026), the same full pass would also require roughly \$3.52 in API cost. We therefore use FACTUAL as the default parser in the training pipeline because it offers the more practical trade-off between parse quality, speed, and operating cost for repeated offline parsing.

Parser	Sentence-level quality audit (48 sentences)			Caption-level efficiency audit (100 captions)		
	Obj.	Attr.	Atom	Sec./cap.	Full pass (min)	API cost/pass (\$)
FACTUAL parser	92.9	85.8	86.7	0.56	17.7	0.00
GPT-5.4 mini	89.2	91.3	87.1	3.46	109.1	3.52

Table 8: Sentence-level parser-quality audit and caption-level efficiency audit for choosing the scene-graph parser used by our method. Quality is measured on 48 sampled training sentences from the TemporalBench short-caption pool with a GPT-5.4 judge. *Atom pass* asks whether a parse is usable for the object/attribute factual supervision used by our reward. Efficiency is measured on 100 training captions under the same sentence-by-sentence parsing pipeline and projected to the full 1,891-caption training pool. API cost is estimated from observed token usage and the public GPT-5.4 mini pricing in March 2026 (OpenAI, 2026).

4.2 Training-data provenance and overlap audit

Our default training set is the 1,891-example TemporalBench short-caption split, where each caption is paired with a unique video. Concretely, this pool is drawn from the TemporalBench short-caption subset spanning six source collections: ActivityNet, Charades, COIN, EgoExo4D, Movie Description, and Oops. In the data-scale ablation, the 150% setting adds

946 extra captions recovered from the TemporalBench QA splits (70 from ‘test_short_qa’ and 876 from ‘test_long_qa’), which introduces 866 additional videos and one extra source collection, FineGym, but still stays inside the TemporalBench video pool.

Training pool	Caps.	Videos	TempComp.	TVBench	VMME	MVB	VidHall.	EventHall.	Total
Default 100%	1891	1891	0/410	0/2112	0/900	0/3851	0/1072	0/397	0/8193
Scaled 150%	2837	2757	0/410	0/2112	0/900	0/3851	0/1072	0/397	0/8193

Table 9: Train–evaluation overlap audit for the TemporalBench-based training pools. We report exact normalized filename overlap against the official benchmark indices or dataset manifests. ‘Total’ is computed against the union of all six benchmark video sets, which contains 8,193 unique videos.

To address the concern that gains may come from benchmark overlap or data leakage, we explicitly audit exact video reuse against the released benchmark indices or official dataset manifests for TempCompass (Liu et al., 2024), TVBench (Cores et al., 2024), Video-MME (Fu et al., 2025), MVBench (Li et al., 2024), VideoHalluciner (Wang et al., 2024b), and EventHallusion (Zhang et al., 2024a). Table 9 reports the result. For both the default 100% training pool and the enlarged 150% pool, we find zero exact video overlap with all six benchmarks after normalizing video filenames. The union of these benchmark video sets contains 8,193 unique evaluation videos, so the observed gains on the audited benchmarks cannot be explained by direct reuse of evaluation clips in training. We likewise use the remaining held-out benchmarks strictly for evaluation and never incorporate their annotations into the training set.

5 Additional Experimental Results

5.1 Additional VDC Caption-Generation Results

Because the motivation of our method originates from video captioning, we additionally evaluate on the VDC caption-generation benchmark using the default Qwen3-VL-8B backbone. Table 10 reports the GPT-4o-mini judged score and binary accuracy for five caption types. Our method yields a small but consistent gain over the base model: the average judged score improves from 2.274 to 2.319, and the average binary accuracy improves from 43.97 to 44.82. The gain is consistent across short, main-object, detailed, and camera-focused captions, while the background subset is essentially tied. We view this result as useful supporting evidence that the same structured reward improving grounded QA and hallucination-sensitive evaluation also transfers back to the caption-generation setting that originally motivated the method.

Caption Type	Score ↑		Acc. ↑	
	Base	+OURS	Base	+OURS
Short	2.274	2.348	44.04	45.49
Main object	2.345	2.404	45.50	46.51
Detailed	2.512	2.553	48.74	49.50
Camera	2.125	2.171	40.81	41.85
Background	2.116	2.117	40.74	40.73
Average	2.274	2.319	43.97	44.82

Table 10: VDC caption-generation results on the default Qwen3-VL-8B backbone. We report the GPT-4o-mini judged mean score (0–5) and binary accuracy for each caption type.

5.2 Additional Scaling Results

Table 11 gives the complete per-benchmark numbers behind the data-scaling curve in Figure 4. The main pattern is that increasing the RL-stage caption pool used by our method

improves all three benchmark groups, with the clearest absolute gains on the hallucination-oriented evaluation. The temporal and conventional gains are smaller but remain positive, which is consistent with the view that extra training data mostly helps factual support and temporal grounding rather than generic answer-format adaptation.

Table 12 reports the full model-scaling results that underlie the backbone-size trend in Figure 4. The 32B backbone is already substantially stronger before RL, so the post-training margin becomes smaller, but the improvement direction remains stable on the available benchmarks. This is the expected scaling behavior for a post-training method that mainly repairs structured consistency rather than changing the underlying visual encoder capacity.

Scale	N	Temporal Understanding			Conventional Video Understanding			Hallucination Examination		
		Temp Compass	TVBench	Avg.	Video-MME	MVBench	Avg.	Video Halluciner	Event Hallusion	Avg.
Baseline	–	74.1	48.6	61.35 $\pm$ 0.00	56.7	65.8	61.25 $\pm$ 0.00	55.7	66.8	61.25 $\pm$ 0.00
10%	189	74.2	48.7	61.45 $\pm$ 0.10	56.9	65.9	61.40 $\pm$ 0.15	55.8	72.1	63.95 $\pm$ 2.70
25%	473	74.4	48.7	61.55 $\pm$ 0.20	56.9	66.3	61.60 $\pm$ 0.35	55.9	72.1	64.00 $\pm$ 2.75
50%	946	74.5	49.1	61.80 $\pm$ 0.45	56.9	66.6	61.75 $\pm$ 0.50	56.2	72.8	64.50 $\pm$ 3.25
100%	1891	75.1	49.2	62.15 $\pm$ 0.80	57.4	66.7	62.05 $\pm$ 0.80	56.8	73.4	65.90 $\pm$ 4.65
150%	2837	75.2	49.1	62.15 $\pm$ 0.80	57.5	66.8	62.15 $\pm$ 0.90	57.0	75.2	66.10 $\pm$ 4.85

Table 11: Ablation study on training data scale. Each group-wise Avg. is the arithmetic mean of the two benchmarks in that group, and the final Avg. is the arithmetic mean over all six benchmarks. Subscripts indicate the change relative to Baseline.

Category	Benchmark	Base			Ours		
		4B	8B	32B	4B	8B	32B
Temporal	TempCompass	71.0	74.1	77.2	73.0 $\pm$ 2.0	75.1 $\pm$ 1.0	77.3 $\pm$ 0.1
	TVBench	41.7	47.0	51.3	47.1 $\pm$ 5.4	48.0 $\pm$ 1.0	51.4 $\pm$ 0.1
Conventional	Video-MME	53.6	56.7	60.1	54.4 $\pm$ 0.8	57.4 $\pm$ 0.7	60.3 $\pm$ 0.2
	MVBench	62.7	65.8	69.5	64.2 $\pm$ 1.5	66.7 $\pm$ 0.9	69.8 $\pm$ 0.3
Hallucination	VideoHalluciner	51.4	55.7	62.0	55.8 $\pm$ 4.4	56.8 $\pm$ 1.1	62.5 $\pm$ 0.5
	EventHallusion	68.4	66.8	77.0	72.0 $\pm$ 3.6	73.4 $\pm$ 6.6	78.0 $\pm$ 1.0

Table 12: Model-size ablations for Qwen3-VL, grouped into temporal understanding, conventional video understanding, and hallucination-oriented evaluation.

5.3 Additional Ablation Tables

Table 13 provides the numeric version of the reward-component ablation summarized in Figure 3. The main trend matches the grouped plot: the full components remains strongest overall, and each reward branch contributes a different part of the final gain profile.

Setting	Temporal Understanding			Conventional Video Understanding			Hallucination Examination		
	TempCompass	TVBench	Avg.	VideoMME	MVBench	Avg.	VideoHalluciner	EventHallusion	Avg.
Ours	75.1	48.0	61.55	57.4	66.7	62.05	56.8	73.4	65.10
Qwen3-VL-8B (Base)	74.1	47.0	60.55 $\pm$ 1.00	56.7	65.8	61.25 $\pm$ 0.80	55.7	66.8	61.25 $\pm$ 3.85
w/o SG Reward	74.6	47.4	61.00 $\pm$ 0.55	57.0	66.1	61.55 $\pm$ 0.50	56.2	71.3	63.75 $\pm$ 1.35
w/o Temp Reward	74.2	47.2	60.70 $\pm$ 0.85	57.1	66.5	61.80 $\pm$ 0.25	56.6	71.5	64.05 $\pm$ 1.05
w/o VQA Reward	74.8	47.7	61.25 $\pm$ 0.30	57.2	65.9	61.55 $\pm$ 0.50	56.0	69.9	62.95 $\pm$ 2.15

Table 13: Ablation studies on our training strategy. SG refers to the scene graph reward. Benchmarks are grouped into temporal understanding, conventional video understanding, and hallucination.

5.4 Additional Qualitative Study

To evaluate the results of our method, we provide a compact qualitative comparison based on representative examples drawn from the saved model outputs on benchmark videos. We highlight cases in which the base model fails despite clear visual evidence—for example, on object identity, procedural steps, temporal order, or scene transitions—while our method corrects the same error for the identical video input. Figures 6 and 5 present two long-form captioning examples and four QA-style examples, respectively.

Qualitative VQA Study



Temporal QA
TempCompass_YorN

Question
Is passing the ball to each other and shooting the basketball the correct order of events?

GT
yes

Base
no

Ours
yes



Procedural MCQ
Video-MME

Question
What was the next step taken by the operator after filling the bottle with oil? Options: A. Applied glue to cork. / B. Added blue food dye. / C. Added 1/3 water. / D. Put the cork on the bottle.

GT
A. Applied glue to cork.

Base
C. Added 1/3 water.

Ours
A. Applied glue to cork.



Open-ended VQA
Vid4alluc_STH

Question
Does the scene change? If yes, identify the source and destination locations.

GT
Scene change: Yes, Locations: from in a room to in space.

Base
Scene change: Yes, Locations: from a room with a heater to a space scene with asteroids and a spaceship.

Ours
Scene change: Yes, Locations: from indoors to space



Fine-grained MCQ
MVBench

Question
What happened before the person closed the closet/cabinet?

GT
Took the sandwich.

Base
(B) Put down the phone/camera.

Ours
(C) Took the sandwich.

Figure 5: Compact qualitative QA and verification comparison. The cases cover temporal QA, procedural MCQ, fine-grained MCQ, and open-ended scene-change reasoning. Across all four examples, the base model makes a locally plausible but grounded error, whereas our method recovers the correct answer with less ambiguity.

Qualitative Captioning Study



Captioning Case A
TemporalBench

Question
Write a detailed caption describing the container, the pill-taking action, and the subsequent drink.

GT
Ground truth sequence: opens a white container, takes a pill, closes the container, reaches for a glass, drinks from it, and puts the glass back down.

Base
The man picks up a small white bottle, opens it, pours liquid into a clear glass, drinks it, and places the glass back on the table.

Ours
A man takes a pill with his right hand, then picks up a glass, drinks from it, and puts the glass back on the table.



Captioning Case B
TemporalBench

Question
Write a detailed caption that describes the human interactions, body placement, and event sequence in the clip.

GT
Ground truth highlights a structured interaction: arm around neck, hand on shoulder, kissing, lifting, legs wrapping, skirt tugging, spinning, and tumbling to the ground.

Base
A man lifts a woman into his arms, spins her around, and gently sets her down while they smile and embrace on the beach.

Ours
The man lifts the woman up with his right hand and holds her waist with his left hand. The woman puts her right arm around his head and her left hand on his shoulder. He spins with her in his arms, her dress flaring around her legs, before they fall to the ground.

Figure 6: Compact qualitative captioning comparison. The selected cases are drawn from the stored TemporalBench short-caption outputs and highlight failures where the base model produces a plausible but factually wrong long caption, while our method restores the missing action or interaction details.