SAW: SINK-AWARE WEIGHTING FOR EFFECTIVE VIDEO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video reasoning requires linking evidence across time, yet providing frames spanning a video does not ensure that a multimodal large language model (MLLM) uses that evidence. We investigate frame-level attention sink, where response tokens concentrate attention on a small subset of frames. Our analysis shows that hallucinated tokens exhibit lower frame-attention entropy and higher peak attention than non-hallucinated tokens. We further derive an entropy-dependent bound on the single-frame approximation error of an attention head's visual contribution. Motivated by these findings, we propose Sink-Aware Weighting (SAW), which reduces the training influence of tokens with relatively concentrated frame attention. SAW converts frame-attention entropy into bounded token weights for supervised fine-tuning losses and reinforcement-learning advantages, while preserving the backbone architecture. With an 8B backbone, SAW achieves 64.99% average accuracy across five video question-answering benchmarks, 2.79 percentage points above the strongest evaluated open-weight baseline with complete results and comparable to GPT-4o (64.86%). On VideoHallucer, it improves pair-level accuracy by 7.60 percentage points over the strongest evaluated baseline. These results support attention-based reweighting for video reasoning and hallucination reduction. Data, code, and models will be made publicly available.

1 INTRODUCTION

Video-capable multimodal large language models (MLLMs) support event description and question answering (Maaz et al., 2024; Bai et al., 2025b). Building on language-model reasoning advances (DeepSeek-AI et al., 2025), chain-of-thought (CoT) supervision and reinforcement learning (RL) have improved temporal and multi-step video reasoning (Feng et al., 2025; Li et al., 2025c). Subsequent work broadens task coverage through unified image–video training (Feng et al., 2026) and improves efficiency by reasoning selectively (Liu et al., 2026).

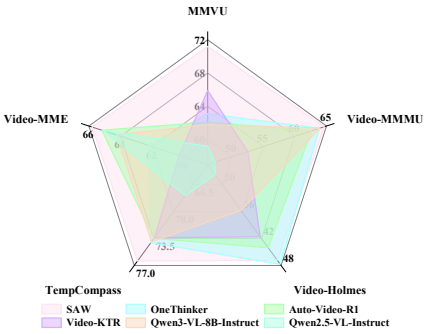

Figure 1: SAW and representative baselines on five video question-answering benchmarks.

Standard supervised fine-tuning (SFT) weights response-token losses uniformly, while outcome-supervised RL with group relative policy optimization (GRPO) (Shao et al., 2024) shares one advantage across a response. These objectives do not explicitly distinguish tokens by their use of video evidence. Video-KTR (Wang et al., 2026) introduces selective token-level RL. We investigate whether frame-attention concentration can provide a continuous weighting signal shared by SFT and RL.

Building on visual attention sinks (Kang et al., 2025) and temporal attention collapse (Sun et al., 2026), we investigate *frame-level attention sink* at reasoning-token granularity. Hallucinated tokens on VideoHallucer (Wang et al., 2024b) exhibit lower frame-attention entropy and higher peak attention (Figure 2a–b). A complementary head-level bound links entropy to single-frame approximation

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.

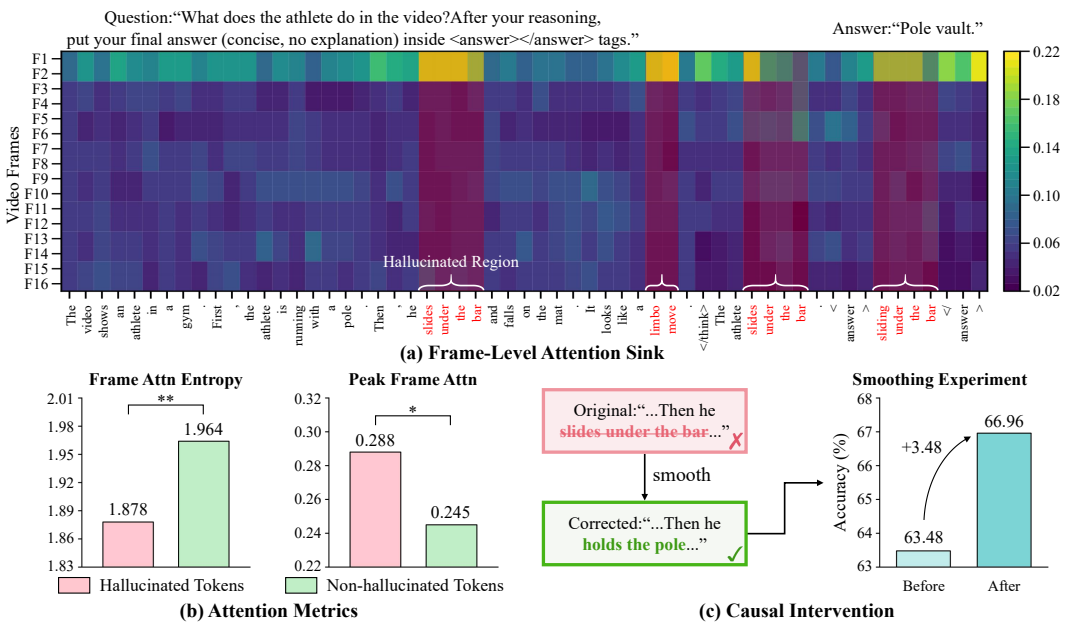

Figure 2: **Frame-attention concentration and hallucinated reasoning.** (a) Per-token attention heatmap over a 16-frame clip; hallucinated tokens (red) concentrate on a narrow band of early frames (F1–F2). (b) Frame attention entropy and peak attention on hallucinated vs. non-hallucinated tokens; significance: $^{*}p<0.05$, $^{**}p<0.01$. (c) Hallucination smoothing: replacing the hallucinated span (red) with a corrected, visually-grounded span (green) and re-feeding the prefix.

error (Appendix A). Separately, correcting hallucinated spans before continuing generation improves answer accuracy by 3.48 percentage points (Figure 2c). These findings motivate attention-based weighting and corrected supervision.

We propose **Sink-Aware Weighting (SAW)**, which maps frame-attention entropy to positive, bounded token weights. A shared *Sink Modulator* scales cross-entropy losses during SFT and group-relative advantages during RL without changing the backbone architecture. We also construct SAW-CoT by regenerating, correcting, and filtering Video-R1 traces (Feng et al., 2025) to improve SFT supervision. SAW uses relatively concentrated frame attention as a proxy for weak visual evidence use. Smaller weights directly attenuate these tokens' gradient contributions in both SFT and RL.

With an 8B backbone, SAW achieves 64.99% average accuracy across five video question-answering benchmarks (Figure 1), 2.79 percentage points above the strongest evaluated open-weight baseline with complete results. On VideoHallucer, it reaches 70.26% pair-level accuracy, a 7.60-point gain over the strongest evaluated baseline.

Our contributions are threefold:

- We identify a token-level association between hallucination and frame-attention concentration, and provide a complementary entropy-dependent approximation bound that motivates concentration-aware training.
- We unify entropy-based weighting across SFT and RL. Positive, bounded weights reduce the influence of relatively concentrated tokens while retaining learning signals for steps that legitimately focus on few frames.
- We demonstrate the complementary value of supervision refinement and attention-based optimization. Ablations show gains from RL weighting with the SFT checkpoint and corpus held fixed, while the complete pipeline improves both video question answering and hallucination evaluation.

2 Related Work

Video understanding and reasoning. Building on visual instruction tuning (Liu et al., 2023; Chen et al., 2024), video-capable MLLMs (Li et al., 2025b; Maaz et al., 2024; Lin et al., 2024; Wang et al., 2024a; Bai et al., 2025b) connect visual representations to language decoders and are evaluated on benchmarks such as MVBench and EgoSchema (Li et al., 2024; Mangalam et al., 2023). Recent methods transfer chain-of-thought prompting (Wei et al., 2022) and RL-based reasoning (Shao et al., 2024; DeepSeek-AI et al., 2025) to video. Video-R1 (Feng et al., 2025) introduces video reasoning data and temporally aware GRPO; VideoChat-R1 (Li et al., 2025c) emphasizes spatiotemporal perception; OneThinker (Feng et al., 2026) unifies image and video reasoning; VideoAuto-R1 (Liu et al., 2026) reasons selectively using initial-answer confidence. Video-KTR (Wang et al., 2026) performs token-level RL using visual dependence, temporal sensitivity, and predictive entropy. SAW instead derives continuous, bounded token weights from frame-attention concentration and applies them during both SFT and RL.

Hallucination evaluation and mitigation. POPE and HallusionBench study image-grounded hallucination (Li et al., 2023; Guan et al., 2024); VideoHallucer, VidHalluc, and EventHallusion extend evaluation to video (Wang et al., 2024b; Li et al., 2025a; Zhang et al., 2024). Visual contrastive decoding compares original and distorted visual inputs (Leng et al., 2024); temporal contrastive decoding disrupts temporal cues (Zhang et al., 2024), while MACD constructs targeted counterfactual inputs (Xiao & Zhou, 2026). OPERA uses an attention-based over-trust penalty and rollback (Huang et al., 2024), and Woodpecker corrects generated hallucinations (Yin et al., 2024). DINO-HEAL reweights visual features using DINOv2 saliency at inference (Li et al., 2025a). For training, RLHF-V uses fine-grained human corrections (Yu et al., 2024), while HSA-DPO uses AI feedback and hallucination severity (Xiao et al., 2025). MASH-VLM modifies attention and positional encoding (Bae et al., 2025); ViSSRes learns residual corrections to video representations with a frozen backbone (Gao et al., 2026). SAW-CoT similarly uses correction for supervision, while SAW weights training updates using attention statistics.

Attention sinks and visual grounding. StreamingLLM studies attention sinks in language models (Xiao et al., 2024); registers address high-norm background tokens in vision Transformers (Darcet et al., 2023). VAR identifies visual sink tokens and redistributes their attention at inference time (Kang et al., 2025). Most directly related, SmartSight links temporal attention collapse to video hallucination and selects among candidate responses using attention statistics (Sun et al., 2026). We extend this line to token-level training: SAW reweights SFT losses and RL advantages rather than selecting responses or directly redistributing attention. Its concentration score remains a proxy, since localized events may legitimately require attention to few frames.

3 Motivation: Frame Attention and Reasoning Quality

SAW is motivated by two questions: how does frame attention relate to hallucinated reasoning, and can improving intermediate claims help the final answer? We study Qwen3-VL-Thinking-8B (Bai et al., 2025a) on 1,563 samples from VideoHallucer (Wang et al., 2024b), whose adversarial question pairs support hallucination analysis. For each response, we extract the language decoder's last-layer attention and use Gemini-3.1-Pro (Google DeepMind, 2026) to identify hallucinated tokens. We measure *frame attention entropy* H_t (Equation (2)), which summarizes how broadly each token attends across frames, and *peak frame attention*, the maximum attention mass on a single frame.

Finding 1: Hallucinated tokens attend more narrowly across the video. Figure 2(a) illustrates the gap between available video context and concentrated frame attention: tokens in the hallucinated span ("slides under the bar") attend mainly to early frames (F1–F2), while non-hallucinated tokens attend more broadly. Across the analyzed samples, hallucinated tokens exhibit lower frame attention entropy (1.878 vs. 1.964, $p<0.01$) and higher peak frame attention (0.288 vs. 0.245, $p<0.05$; Figure 2b). Proposition 1 in Appendix A complements this observation: lower entropy tightens a head-level dominant-frame approximation bound when visual attention mass and frame-value dispersion are controlled (Equation (10)). Together, these results motivate entropy as a concentration proxy for the Sink Modulator (Section 4.2). Entropy alone does not determine visual correctness, since localized events may appropriately require attention to few frames.

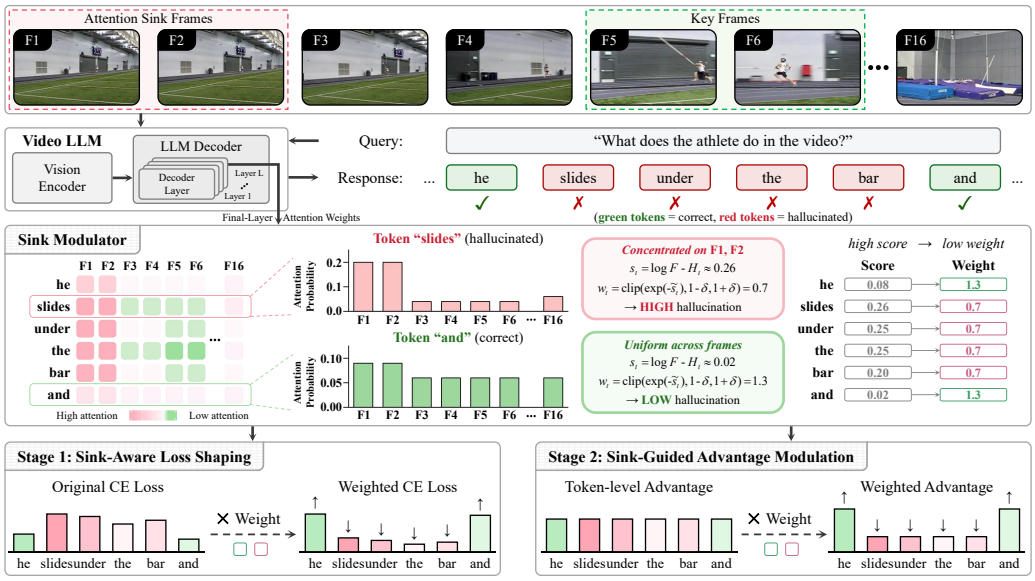

Figure 3: **SAW overview. Top:** the Sink Modulator reads the language decoder's last-layer attention and assigns each response token a weight from its frame-attention entropy—lower for relatively concentrated attention (red), higher for relatively distributed attention (green). **Bottom-left:** In the SFT stage, the weights are applied to the per-token cross-entropy (CE) loss. **Bottom-right:** In the RL stage, the weights modulate per-token GRPO advantages across all responses in the default setting.

Finding 2: Correcting intermediate claims improves final answers. We use the same MLLM judge to rewrite each detected hallucinated span with a minimum-edit, visually-grounded correction, then continue generation from the revised prefix. This *hallucination smoothing* increases accuracy from 63.48% to 66.96%, a gain of 3.48 percentage points (Figure 2c). Thus, correcting intermediate claims can improve subsequent predictions. The intervention changes the text prefix; it does not manipulate attention or isolate attention sink as a cause of hallucination.

These two findings motivate complementary changes to supervision and optimization. The token-level association suggests a concentration-based signal for weighting training contributions. The prefix-correction result motivates revising unsupported intermediate claims in the supervision. SAW-CoT changes the target reasoning text, while a shared weighting rule adjusts each token's contribution during SFT and RL. The two components therefore address the content of supervision and its influence on learning, respectively.

4 Sink-Aware Weighting

Overview. SAW pairs corrected supervision with concentration-aware optimization, following Section 3 (Figure 3). We train on SAW-CoT with Sink-Aware Loss Shaping (SALS), then continue from this checkpoint with Sink-Guided Advantage Modulation (SGAM). A shared Sink Modulator converts frame-attention concentration into detached, positive, bounded token weights for both stages.

4.1 SAW-CoT: Reannotating Video Reasoning

We construct SAW-CoT by reannotating Video-R1 (Feng et al., 2025) reasoning traces for SALS (Section 4.3). Retained examples preserve their source videos, questions, and reference answers.

Figure 4 summarizes generation, detection, correction, and filtering. Gemini-3.1-Pro (Google DeepMind, 2026) generates reasoning traces, flags spans unsupported by the video, and minimally corrects them. We retain traces whose final answers agree with the reference, using rule-based

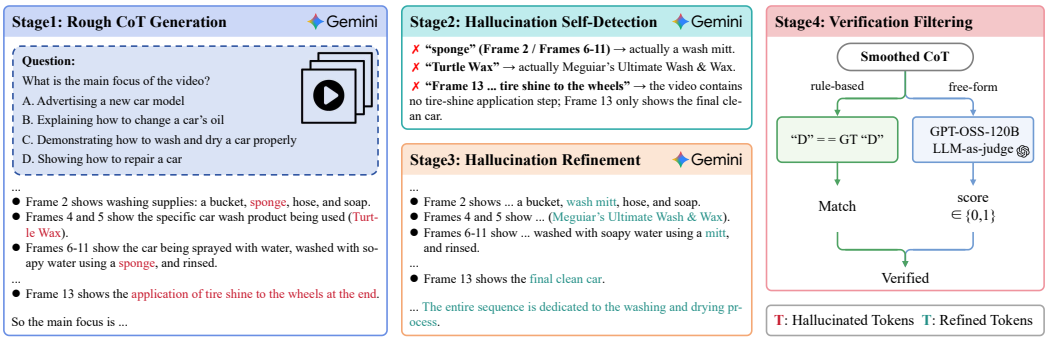

Figure 4: **SAW-CoT construction.** Reasoning traces are regenerated for Video-R1 examples, checked for unsupported spans, minimally corrected, and filtered by final-answer agreement with the reference. Stage 4 uses rule-based matching for closed-form answers and GPT-OSS-120B for free-form answers.

matching for closed-form questions and GPT-OSS-120B for free-form questions (Appendix B). This filter does not certify every intermediate claim.

4.2 Sink Modulator

Frame-attention aggregation. Let $\mathbf{x} = (x_1, \ldots, x_P)$ be a multimodal prompt with F frames and $\mathbf{y} = (y_1, \ldots, y_N)$ a response. Let $A_{q,j}$ denote extracted last-layer attention from decoder query q to prompt token j, and $q = \kappa(t)$ the query aligned with response token y_t. Special tokens mark frame boundaries, defining the token positions $\mathcal{F}_f$ of frame f. Its attention mass is

$$a_{t,f} = \sum_{j \in \mathcal{F}_f} A_{\kappa(t),j}. \tag{1}$$

Entropy-based scoring. The head-level bound discussed in Section 3 motivates using frame-attention entropy as a concentration proxy. For positive visual attention mass $m_t = \sum_f a_{t,f}$, we normalize over frames and compute the entropy deficit:

$$p_{t,f} = \frac{a_{t,f}}{m_t}, \qquad H_t = -\sum_{f=1}^{F} p_{t,f} \log p_{t,f}, \qquad s_t = \log F - H_t. \tag{2}$$

Here $\log$ is natural and $0 \log 0 = 0$. The score ranges from 0 for uniform attention to $\log F$ for single-frame concentration. Since normalization removes m_t, H_t measures how visual attention is distributed, not how much attention the video receives.

Weight calibration. We standardize scores within each response, $z_t = (s_t - \mu_s)/\sigma_s$ for $\sigma_s > 0$, and map them to bounded weights:

$$w_t = \mathrm{sg}\big[\mathrm{clip}\big(e^{-z_t},\, 1-\delta,\, 1+\delta\big)\big], \qquad 0 < \delta < 1, \tag{3}$$

where μ_s, σ_s are the response-level mean and standard deviation, and sg stops gradients. Above-average concentration gives $w_t < 1$, while below-average concentration gives $w_t > 1$. We use $\delta = 0.2$ for SFT and $\delta = 0.3$ for RL, giving weights in $[0.8, 1.2]$ and $[0.7, 1.3]$. Figure 5 evaluates the clipping radius and probe layer. Since F is fixed within a response, z_t is also the negative standardized entropy. The weights reflect relative concentration and need not average to one. Detaching them reweights learning without directly maximizing entropy.

Task-dependent frame relevance. Frame relevance depends on the question and reasoning step: local events may require few frames, while temporal comparisons or global understanding may require distributed evidence. Key-frame guidance therefore needs a relevance criterion or selector. SAW requires no target-frame labels. Its positive, bounded weights retain each weighted token's learning contribution and limit modulation, since concentration alone cannot distinguish useful focus from spurious sinks.

4.3 SINK-AWARE LOSS SHAPING

For a target response $\mathbf{y}$ from SAW-CoT, SALS trains policy π_θ by applying w_t from Equation (3) to each token's cross-entropy loss:

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{N}\sum_{t=1}^{N} w_t \log \pi_\theta(y_t \mid \mathbf{x}, y_{<t}). \tag{4}$$

Writing $\ell_t = -\log \pi_\theta(y_t \mid \mathbf{x}, y_{<t})$, detaching the weight gives $\nabla_\theta(w_t \ell_t) = w_t \nabla_\theta \ell_t$. Thus, smaller weights attenuate the update contribution of relatively concentrated tokens (Figure 3, bottom-left). Table 3 compares standard and weighted SFT on SAW-CoT.

4.4 SINK-GUIDED ADVANTAGE MODULATION

Rollouts and rewards. Starting from the SALS checkpoint, we train on the filtered RL corpus described in Appendix C using outcome-supervised GRPO (Shao et al., 2024; DeepSeek-AI et al., 2025). For each prompt $\mathbf{x}$, the rollout policy $\pi_{\theta_{\text{old}}}$ generates G responses $\{\mathbf{y}^{(i)}\}_{i=1}^{G}$ of lengths N_i. Each receives reward $R^{(i)}$, combining answer correctness and format (Appendix C). With group mean μ_R and standard deviation $\sigma_R > 0$, the sequence advantage is

$$\hat{A}^{(i)} = \frac{R^{(i)} - \mu_R}{\sigma_R}. \tag{5}$$

Let $\rho_t^{(i)} = \pi_\theta(y_t^{(i)} \mid \mathbf{x}, y_{<t}^{(i)}) / \pi_{\theta_{\text{old}}}(y_t^{(i)} \mid \mathbf{x}, y_{<t}^{(i)})$. With clipping radius ϵ, the per-token policy loss (Schulman et al., 2017) is

$$\ell_{\text{clip}}(\rho, \hat{A}) = -\min\Big\{\rho\hat{A},\ \text{clip}(\rho, 1-\epsilon, 1+\epsilon)\hat{A}\Big\}. \tag{6}$$

Standard GRPO averages this loss over tokens within each response and then over responses. We omit the KL-to-reference penalty.

Advantage modulation. SGAM uses rollout attention to compute response-normalized weights via Equation (3), applying them to all responses regardless of final-answer correctness. For positive w, $\ell_{\text{clip}}(\rho, w\hat{A}) = w\ell_{\text{clip}}(\rho, \hat{A})$, so modulating the advantage is exactly equivalent to weighting the clipped loss:

$$\begin{aligned} \hat{A}_t'^{(i)} &= w_t^{(i)} \hat{A}^{(i)}, \\ \mathcal{L}_{\text{SGAM}} &= \frac{1}{G}\sum_{i=1}^{G} \frac{1}{N_i}\sum_{t=1}^{N_i} w_t^{(i)} \ell_{\text{clip}}\Big(\rho_t^{(i)}, \hat{A}^{(i)}\Big). \end{aligned} \tag{7}$$

Equations (4) and (7) share the same update rule: a detached weight scales each token's base-loss gradient. Positive weights preserve advantage signs and clipping branches, while weights below 1 attenuate both positive and negative contributions. SGAM supplies no token-level correctness reward. Table 3 compares GRPO and SGAM with the same initialization and RL data, and Table 4 tests which responses to weight.

5 EXPERIMENTS

5.1 SETUP

Benchmarks. We evaluate hallucination on VideoHallucer (Wang et al., 2024b) (intrinsic and extrinsic, 1,952 samples). Five further benchmarks cover expert-level reasoning (MMVU (Zhao et al., 2025)), knowledge acquisition (Video-MMMU (Hu et al., 2025)), multi-step reasoning (Video-Holmes (Cheng et al., 2025)), temporal understanding (TempCompass (Liu et al., 2024)), and general video understanding (Video-MME (Fu et al., 2025)).

Table 1: **Main results on video question-answering benchmarks** (accuracy, %). Avg. is the mean over all five benchmarks. – denotes unreported entries. Best open-weight result in **bold**, second best underlined. ‡: evaluated by us under our 32-frame protocol.

Model	Size	Frames	Video reasoning			Video perception		Avg.
			MMVU	Video-MMMU	Video-Holmes	Temp-Compass	Video-MME	
CLOSED-SOURCE								
GPT-4o	–	–	75.40	61.20	42.00	73.80	71.90	64.86
Gemini-2.5-Pro	–	–	–	83.60	45.00	–	84.30	–
QWEN2.5-VL-BASED								
Qwen2.5-VL-Instruct	7B	32	59.20	47.40	27.80	67.90	65.10	53.48
Video-R1	7B	32	63.80	52.30	–	73.20	59.30	–
Video-KTR	7B	32	65.90	52.60	41.60	73.40	60.30	58.76
VideoChat-R1	7B	–	–	46.40	33.00	–	60.00	–
QWEN3-VL-BASED								
Qwen3-VL-Instruct‡	8B	32	61.80	64.00	35.71	73.92	64.26	59.94
OneThinker‡	8B	32	63.10	63.89	**47.74**	74.05	62.22	62.20
Auto-Video-R1‡	8B	32	62.10	62.67	44.04	73.48	65.19	61.10
SAW (ours)	8B	32	**71.20**	**64.67**	46.83	**76.43**	**65.81**	**64.99**

Implementation details. We use Qwen3-VL-8B-Instruct (Bai et al., 2025a) as the backbone and train SAW with SFT on SAW-CoT followed by RL on a curated Video-R1/Auto-Video-R1 corpus (Section 4). Our models and the baselines marked ‡ in Table 1 use 32 uniformly sampled frames and a maximum generation length of 8,192 tokens at evaluation. Closed-source results for GPT-4o and Gemini-2.5-Pro (Comanici et al., 2025) use default API settings, so their input and generation settings differ. Appendix C gives training settings, reward weights, data filtering, and compute resources.

5.2 MAIN RESULTS

Video reasoning and general video understanding. SAW improves over its Qwen3-VL-Instruct backbone on every benchmark in Table 1, covering both reasoning and perception tasks. It achieves the highest average among evaluated open-weight models with complete results, exceeding OneThinker by 2.79 percentage points. SAW leads on most individual benchmarks, while OneThinker remains strongest on Video-Holmes. Its average is comparable to GPT-4o, although their input and generation settings differ. The ablations below examine which training components contribute to these gains.

Hallucination benchmarks. SAW achieves the highest VideoHallucer average, exceeding the strongest evaluated baseline, Auto-Video-R1, by 7.60 percentage points (Table 2). The metric requires both answers in a basic/adversarial question pair to be correct and averages the subtypes equally. Improvements over the backbone span intrinsic and extrinsic hallucination, with a particularly large gain on External Factual. Auto-Video-R1 remains strongest on Semantic Detail. OneThinker's result should be interpreted under our fixed generation budget, which truncates some adversarial responses before a final answer.

5.3 ABLATION STUDIES

We conduct ablations on three general video benchmarks (TempCompass, Video-MME, Video-MMMU) and report the average accuracy for each variant. As in the main experiments (Table 1), all ablations use the full training data for their respective configurations.

Component ablation. Table 3 distinguishes supervision refinement from update weighting. Replacing Video-R1 traces with SAW-CoT improves standard SFT, and SALS provides a further gain on the same corpus. These SFT variants remain below the unadapted backbone on average, so corrected traces alone do not account for the final gains. The direct test of RL weighting compares SGAM with GRPO from the same SALS checkpoint and filtered RL corpus (Appendix C). SGAM

Table 2: **Results on VideoHallucer** (pair-level accuracy, %). All models have 8B parameters and use our 32-frame evaluation protocol. Avg. is the mean over the five subtypes. Best results are **bold**, and second best are underlined.

Model	Intrinsic			Extrinsic		Avg.
	Object relation	Temporal	Semantic detail	External factual	External non-factual	
Qwen3-VL-Instruct	66.50	61.93	76.00	36.50	69.00	61.99
OneThinker	14.00	38.64	18.00	9.50	29.00	21.83
Auto-Video-R1	70.50	60.80	**78.00**	32.50	71.50	62.66
SAW	**71.00**	**71.02**	77.00	**53.30**	**79.00**	**70.26**

Table 3: **Component ablation.** All variants use Qwen3-VL-8B-Instruct. Accuracy (%) and its mean over three benchmarks are reported. SAW-CoT is the reannotated corpus, SALS is sink-aware loss shaping, and SGAM is sink-guided advantage modulation. ✓ and – denote active and inactive components. Best results are **bold**.

Training	SAW-CoT	SALS	SGAM	Temp-Compass	Video-MME	Video-MMMU	Avg.
Backbone	–	–	–	73.92	64.26	64.00	67.39
SFT	–	–	–	70.13	57.59	58.56	62.09
SFT	✓	–	–	73.52	58.81	60.11	64.15
SFT	✓	✓	–	75.32	61.74	61.44	66.17
SFT + GRPO	✓	✓	–	74.29	64.07	62.56	66.97
SFT + SGAM (**SAW**)	✓	✓	✓	**76.43**	**65.81**	**64.67**	**68.97**

improves average accuracy by 2.00 percentage points, supporting a benefit beyond data refinement and initialization. This sequential design measures incremental contributions within the pipeline, without establishing independent component effects.

Modulation scope. Applying SGAM to all responses yields the highest average in Table 4 and is our default. Restricting modulation to correct or incorrect responses gives a lower average. Outcome rewards evaluate final answers, whereas SAW measures token-level attention concentration, which can occur in either response group. The results favor applying these complementary signals without restricting modulation by answer correctness, although the modest differences do not establish statistical significance.

Table 4: **SGAM modulation scope.** Accuracy (%). Avg. is the mean across benchmarks. Best in **bold**.

Scope	TempCompass	Video-MME	Video-MMMU	Avg.
Incorrect only	76.05	**65.86**	63.45	68.45
Correct only	76.31	65.10	64.56	68.66
All responses	**76.43**	65.81	**64.67**	**68.97**

Weight clipping and layer selection. Figure 5 tests the strength and source of the weighting signal. Intermediate clipping ranges perform best, supporting moderate reweighting of the base objective. The selected radii are δ=0.2 for SFT and δ=0.3 for RL (Equation (3)). Deeper probe layers perform better among the tested settings, peaking at layer 35. This supports the final-layer probe and agrees with the deep-layer attention differences analyzed below.

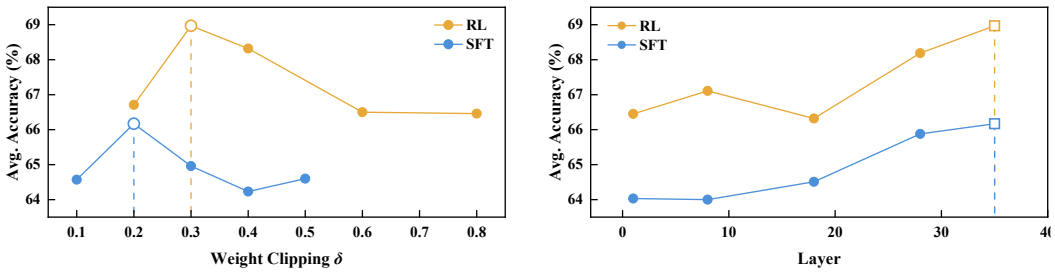

Figure 5: **Hyperparameter ablation. Left:** Weight-clip δ sweep. **Right:** Probe layer selection. Open markers and dashed lines highlight the best tested setting.

5.4 Analysis

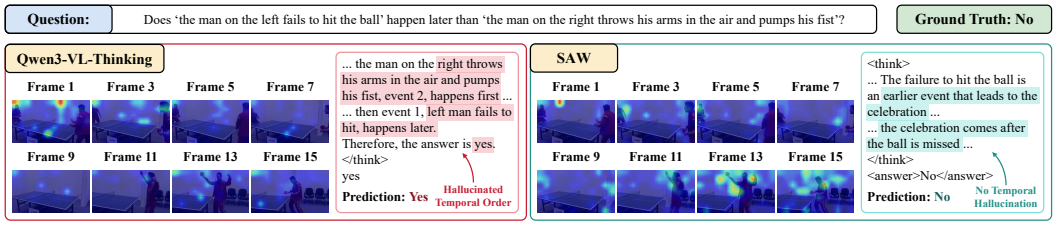

Figure 6: **Qualitative attention comparison** on a temporal-order question from VideoHallucer (Wang et al., 2024b) (ground-truth answer: *No*). **Left:** Qwen3-VL-Thinking-8B baseline concentrates attention on late frames (red overlay). **Right:** SAW distributes attention across both events and answers correctly.

Qualitative attention comparison. Figure 6 contrasts SAW's coverage of both events and correct temporal reasoning with late-frame focus and reversed event order in Qwen3-VL-Thinking-8B. Since SAW starts from Qwen3-VL-8B-Instruct, the example illustrates different model behaviors rather than a controlled training effect.

Layer-wise frame attention. On 200 VideoHallucer samples, SAW maintains higher frame-attention entropy in deep layers, while early-layer patterns are similar across models sharing the Qwen3-VL architecture (Figure 7 and Appendix D). This agrees with the concentration-aware training motivation, although attention differences alone do not establish a causal explanation for the accuracy gains.

Answer accuracy and reasoning quality. Cases 3–4 in Appendix E give correct final answers but include claims that conflict with the paired temporal annotations. These cases motivate evaluating the support for intermediate reasoning separately from answer correctness.

6 Conclusion

SAW reweights SFT losses and GRPO advantages through frame-attention entropy. With an 8B backbone, it achieves 64.99% average accuracy across five video QA benchmarks and 70.26% pair-level accuracy on VideoHallucer. These results support attention-based training for video reasoning.

Limitations. SAW uses frame-attention concentration as a proxy for token weighting. Localized events may legitimately require narrow attention, so the statistic alone cannot determine whether a token uses the right visual evidence. Our evaluation focuses on one 8B backbone, leaving transfer across architectures and scales for further study. SAW-CoT relies on model-generated corrections and answer-based filtering; retained traces may still contain unsupported intermediate claims.

AI USE STATEMENT

Generative AI tools were used to construct and evaluate reasoning traces: Gemini-3.1-Pro generated traces, identified unsupported spans, and proposed corrections, while GPT-OSS-120B assessed answer equivalence for free-form questions (Section 4.1 and Appendix B). Gemini-3.1-Pro also supported the token-level hallucination analysis and prefix-correction experiment in Section 3. OpenAI Codex assisted with language editing, identifying overstatements, interpreting reported results, checking and revising the attention-concentration bound and proof, and adapting the manuscript to the ICLR format. Responsibility for the scientific claims, experimental results, and final manuscript remains with the authors.

ETHICS STATEMENT

This work studies hallucination in video language models using existing video QA data. Improved benchmark performance does not establish reliability in safety-critical applications, and model-generated explanations may remain unsupported or biased. Regenerated and corrected reasoning traces may inherit errors from the source data and automated judges. Any release of derived annotations or models should respect the source datasets' licenses, consent conditions, and privacy constraints.

REPRODUCIBILITY STATEMENT

Section 4 defines the attention statistics, weight calibration, and SFT/RL objectives. Appendix A states the assumptions and proof of the attention-concentration bound. Section 4.1 and Appendix B describe SAW-CoT construction and its prompts, while Appendix C reports the training configuration, evaluation protocol, and RL data filtering. Additional analyses and qualitative examples appear in Appendices D and E. Code and models will be made publicly available.

REFERENCES

Kyungho Bae, Jinhyung Kim, Sihaeng Lee, Soonyoung Lee, Gunhee Lee, and Jinwoo Choi. MASH-VLM: Mitigating action-scene hallucination in video-LLMs through disentangled spatial-temporal representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13744–13753, 2025.

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, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-Holmes: Can MLLM think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025. URL https://arxiv.org/abs/2505.21374.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. URL https://arxiv.org/abs/2507.06261.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 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.

Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, Yan Feng, Peng Pei, Xunliang Cai, and Xiangyu Yue. OneThinker: All-in-one reasoning model for image and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Yuansheng Gao, Jinman Zhao, Tong Zhang, Xingguo Xu, Wenbin Xing, Han Bao, Zonghui Wang, and Wenzhi Chen. Enhancing video representations with spatiotemporal-semantic residual to mitigate hallucinations in video large multimodal models. *arXiv preprint arXiv:2601.22574*, 2026. URL https://arxiv.org/abs/2601.22574v2.

Google DeepMind. Gemini 3.1 Pro model card. Model card, February 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed: 2026-09-26.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. HallusionBench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-MMMU: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025. URL https://arxiv.org/abs/2501.13826.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7uDI7w5RQA.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. VidHalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. VideoChat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102, 2025b.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. VideoChat-R1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025c.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Shuming Liu, Mingchen Zhuge, Changsheng Zhao, Jun Chen, Lemeng Wu, Zechun Liu, Chenchen Zhu, Zhipeng Cai, Chong Zhou, Haozhe Liu, et al. VideoAuto-R1: Video auto reasoning via thinking once, answering twice. *arXiv preprint arXiv:2601.05175*, 2026.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024. doi: 10.18653/v1/2024.findings-acl.517. URL https://aclanthology.org/2024.findings-acl.517/.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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. *arXiv preprint arXiv:2402.03300*, 2024.

Yiming Sun, Mi Zhang, Feifei Li, Geng Hong, and Min Yang. SmartSight: Mitigating hallucination in video-LLMs without compromising video understanding via temporal attention collapse. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(11):9251–9259, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.03762.

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. VideoHallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024b.

Ziyue Wang, Sheng Jin, Zhongrong Zuo, Jiawei Wu, Han Qiu, Qi She, Hao Zhang, and Xudong Jiang. Video-KTR: Reinforcing video reasoning via key token attribution. In *International Conference on Learning Representations*, 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Qixin Xiao and Kun Zhou. MACD: Model-aware contrastive decoding via counterfactual data. *arXiv preprint arXiv:2602.01740*, 2026. URL https://arxiv.org/abs/2602.01740v3.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. Detecting and mitigating hallucination in large vision language models via fine-grained AI feedback. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (24):25543–25551, 2025.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

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*, 2024.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, Chengye Wang, Ziyao Shangguan, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. MMVU: Measuring expert-level multi-discipline video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8475–8489, 2025. URL https://arxiv.org/abs/2501.12380.

A THEORETICAL ANALYSIS: FRAME-ATTENTION CONCENTRATION

We relate frame-attention concentration to the visual contribution of a single attention head. The analysis bounds approximation by one frame's value centroid, then connects the bound to entropy. It does not model hallucination as a direct consequence of attention concentration.

A.1 ATTENTION AND FRAME-LEVEL VALUES

Fix a decoder query position q and let $\mathcal{C}_q$ contain its causally visible key positions, including prompt and visible response tokens. For a head of dimension d_h, define its attention (Vaswani et al., 2017) by

$$\alpha_{q,j} = \frac{\exp(Q_q K_j^\top / \sqrt{d_h})}{\sum_{k \in \mathcal{C}_q} \exp(Q_q K_k^\top / \sqrt{d_h})}, \qquad j \in \mathcal{C}_q. \tag{8}$$

Thus $\sum_{j \in \mathcal{C}_q} \alpha_{q,j} = 1$. Normalization permits concentration but does not require it; equal logits give uniform attention. We use α for a single head to distinguish it from the extracted statistic A in Section 4.2.

For the frame-position sets $\mathcal{F}_f$ in Equation (1), let $b_{q,f} = \sum_{j \in \mathcal{F}_f} \alpha_{q,j}$, $m_q = \sum_f b_{q,f} > 0$, and $r_{q,f} = b_{q,f}/m_q$. The head's visual contribution is $\mathbf{o}_q^{\text{vid}} = \sum_f b_{q,f} \mathbf{v}_{q,f}$, where

$$\mathbf{v}_{q,f} = \frac{1}{b_{q,f}} \sum_{j \in \mathcal{F}_f} \alpha_{q,j} \mathbf{V}_j$$

is the attention-weighted value centroid for $b_{q,f} > 0$. A zero-mass frame contributes nothing; its centroid may be assigned arbitrarily. The remaining visible positions contribute the non-video term of the head output. All quantities in this analysis are head-specific.

A.2 APPROXIMATION BY THE DOMINANT FRAME

Proposition 1 (Dominant-frame approximation)**.** *For $F \geq 2$, choose $f^* \in \arg\max_f r_{q,f}$ and define $D_q = \max_{f \neq f^*} \|\mathbf{v}_{q,f} - \mathbf{v}_{q,f^*}\|_2$. Then*

$$\left\| \mathbf{o}_q^{\text{vid}} - m_q \mathbf{v}_{q,f^*} \right\|_2 \leq m_q (1 - r_{q,f^*}) D_q. \tag{9}$$

In particular, if $r_{q,f^*} \geq 1 - \eta$ for $\eta \in [0,1]$, the bound is at most $m_q \eta D_q$.

Proof. Using $\sum_f b_{q,f} = m_q$, write

$$\mathbf{o}_q^{\text{vid}} - m_q \mathbf{v}_{q,f^*} = \sum_{f \neq f^*} b_{q,f} (\mathbf{v}_{q,f} - \mathbf{v}_{q,f^*}).$$

The triangle inequality bounds its norm by $D_q \sum_{f \neq f^*} b_{q,f} = m_q (1 - r_{q,f^*}) D_q$. □

To connect this result to Equation (2), let $H_q^{\text{head}} = -\sum_f r_{q,f} \log r_{q,f}$. Since $r_{q,f} \leq r_{q,f^*}$, averaging $-\log r_{q,f} \geq -\log r_{q,f^*}$ over the positive-probability frames gives $H_q^{\text{head}} \geq -\log r_{q,f^*}$. Consequently,

$$\begin{aligned} r_{q,f^*} &\geq e^{-H_q^{\text{head}}}, \\ \left\| \mathbf{o}_q^{\text{vid}} - m_q \mathbf{v}_{q,f^*} \right\|_2 &\leq m_q \left(1 - e^{-H_q^{\text{head}}}\right) D_q. \end{aligned} \tag{10}$$

Low entropy therefore tightens the approximation bound when m_q and D_q are controlled. The statement concerns a head's visual contribution; it does not imply that other frames vanish from the network representation. Their information may remain in other heads, earlier layers, or residual connections, and frame values may already contain information from other frames.

A.3 Connection to Training and Visual Grounding

Let $\mathbf{h}_t^{\text{pred}}$ denote the final hidden state used to predict y_t from $(\mathbf{x}, y_{<t})$. The next-token distribution is

$$\pi_\theta(\cdot \mid \mathbf{x}, y_{<t}) = \text{softmax}(\mathbf{W}_{\text{LM}}\mathbf{h}_t^{\text{pred}}). \tag{11}$$

This state combines information from all heads, layers, and residual pathways. Even maximal concentration in one head need not eliminate relevant video evidence or reduce prediction to a text-only prior. The conditional approximation bound is therefore distinct from a claim about hallucination in the full model.

Together with the token-level association in Section 3, the bound motivates testing concentration as a training proxy. It does not derive the exponential calibration in Equation (3) or establish its optimality. SAW applies the same detached weights to cross-entropy and clipped policy losses (Equations (4) and (7)); the ablations in Section 5.3 evaluate this design empirically.

B SAW-CoT Construction Details

SAW-CoT reannotates Video-R1 (Feng et al., 2025) examples with newly generated and corrected reasoning traces, then filters them by final-answer agreement (Section 4.1). The source video, question, and reference answer for each retained example remain unchanged. We detail the prompts and filtering procedure below.

Stage 1: Rough CoT Generation. We prompt Gemini-3.1-Pro (Google DeepMind, 2026) to generate a chain-of-thought reasoning trace for each video QA sample.

```
You are given a video and a question about it.  Think step by step,
carefully observe the video content, and reason through the answer.

Question:  {question}

Provide your reasoning inside <think></think> tags, then give your
final answer inside <answer></answer> tags.
```

Stage 2: Hallucination Detection. The same MLLM flags spans in the Stage-1 CoT that it judges unsupported by the visual evidence.

```
You are given a video, a question, and a model-generated reasoning
trace.  Your task is to identify any spans in the reasoning that
describe content NOT actually present in the video.

Question:  {question}
Reasoning:  {cot_response}

For each hallucinated span, output it in the format:
[HALLUCINATION] "<exact span>" | REASON: <why it is not supported>

If there are no hallucinations, output:  [NONE]
```

Stage 3: Hallucination Smoothing. For each flagged span, the MLLM proposes a minimum-edit correction based on the video while preserving the reasoning flow.

```
You are given a video, a question, a reasoning trace, and a list of
hallucinated spans identified in the trace.  For each hallucinated
span, rewrite it with the minimum edit that makes it consistent
with the actual video content.  Preserve the surrounding reasoning
structure.

Question:  {question}
Reasoning:  {cot_response}
Hallucinated spans:  {hallucination_list}

Output the corrected reasoning trace in full.
```

Stage 4: Answer Verification. The final answer accompanying the corrected CoT is checked against the reference answer. For closed-form questions (multiple choice, yes/no), we use rule-based exact matching. For free-form questions, we use GPT-OSS-120B as an LLM-as-judge. Samples whose predicted answer disagrees with the ground truth are discarded.

```
You are a judge evaluating whether a model's answer is correct.
Compare the predicted answer to the ground-truth answer.  The
predicted answer does not need to match the ground truth exactly--it
is correct if it conveys the same meaning.

Question:  {question}
Ground-truth answer:  {gt_answer}
Predicted answer:  {pred_answer}

Output ONLY one of:  [CORRECT] or [INCORRECT]
```

C Implementation Details

Training configuration. We use Qwen3-VL-8B-Instruct (Bai et al., 2025a) as the backbone. The SFT stage is trained on SAW-CoT with weight clipping $\delta = 0.2$, entropy scoring, and learning rate 1×10^{-5}. The RL stage is initialized from the SFT checkpoint and trained with weight clipping $\delta = 0.3$, entropy scoring, GRPO clip ratio $\epsilon = 0.2$, learning rate 1×10^{-6}, and group size $G = 8$; the KL-to-reference penalty is omitted. The reward $R^{(i)}$ in Equation (5) weights the accuracy and format terms with coefficients 0.7 and 0.3. All training and evaluation runs are conducted on a server with 32 NVIDIA A100 GPUs. Variation across training seeds and the additional computational cost of attention collection remain unquantified.

Evaluation protocol. For our models and the baselines marked ‡ in the result tables, we use 32 uniformly sampled frames per video and a maximum generation length of 8,192 tokens. Closed-source models, including GPT-4o and Gemini-2.5-Pro (Comanici et al., 2025), follow their default API settings. The fixed generation budget can truncate long reasoning traces, as observed for OneThinker, and should be considered when interpreting the comparisons.

RL training data. The RL corpus combines (video, question, ground-truth answer) triples from the training splits of Video-R1 and Auto-Video-R1. We apply offline filtering using four responses per prompt from the SFT checkpoint and retain prompts with both correct and incorrect answers. This selects prompts with variation in observed correctness, which can provide a useful group-relative learning signal. Four uniformly correct or incorrect responses do not establish that a prompt is intrinsically easy or unlearnable. We use the same filtered corpus for standard GRPO and SGAM.

D Additional Results

Layer-wise frame attention entropy. We plot per-layer frame-attention entropy averaged over 200 VideoHallucer samples for five models sharing the Qwen3-VL architecture (Figure 7). The models exhibit similar entropy in early layers (0–15), followed by a decline from around layer 18. SAW maintains higher entropy than the other models through layers 20–35, indicating more distributed frame

attention on these samples. This observation is consistent with the intended effect of reweighting, but does not establish that the attention change mediates the accuracy gains.

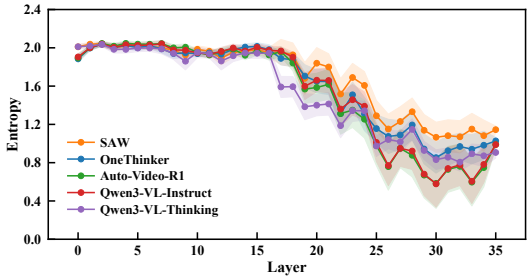

Figure 7: **Per-layer frame attention entropy.** SAW maintains higher entropy in the deep layers (20–35), indicating more distributed frame attention on the analyzed samples.

RL training dynamics. Figure 8 compares RL training dynamics for standard GRPO and SAW. **(a) Policy entropy.** SAW maintains higher policy entropy, indicating a less concentrated token distribution; this quantity is distinct from frame-attention entropy and does not directly measure visual grounding. **(b) Training accuracy.** SAW also achieves higher training accuracy over the plotted trajectory. Together, these trends are consistent with a useful change in optimization behavior. **(c) Average sink weight.** The mean of w_t increases during SAW training. Because the scores are standardized within each response and then exponentiated and clipped, the mean weight depends on the shape of the standardized score distribution. Its increase alone does not demonstrate a decrease in absolute sink severity. The frame-attention entropy analysis above provides a more direct measurement of attention concentration.

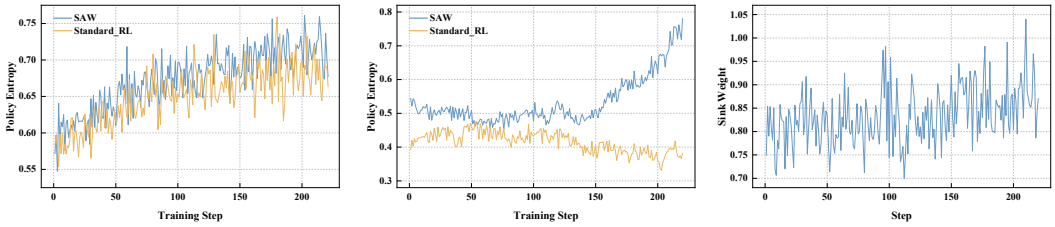

Figure 8: **RL training dynamics.** (a) Policy entropy. (b) Training accuracy. (c) Average sink weight w_t, computed from standardized and clipped per-token scores.

E Qualitative Examples

We present four case studies from VideoHallucer (Wang et al., 2024b) comparing SAW with the Qwen3-VL-Thinking-8B baseline. Each example shows the video frames, question, ground-truth answer, and the full reasoning traces from both models. The figures retain the original green/red span annotations; these are qualitative judgments, not independently verified token-level labels. Correct final answers do not establish that every intermediate claim is grounded.

Event localization. Figures 9–10 test whether a specific event occurs at the beginning of a video. In both cases, the baseline's reasoning drifts from the visual evidence: it incorrectly infers that the target event occurs later in the video despite clear early-frame evidence. SAW, by contrast, correctly grounds its reasoning in the early frames and arrives at the right answer. These examples illustrate errors in temporal localization; the reasoning traces alone do not establish which attention pattern caused them.

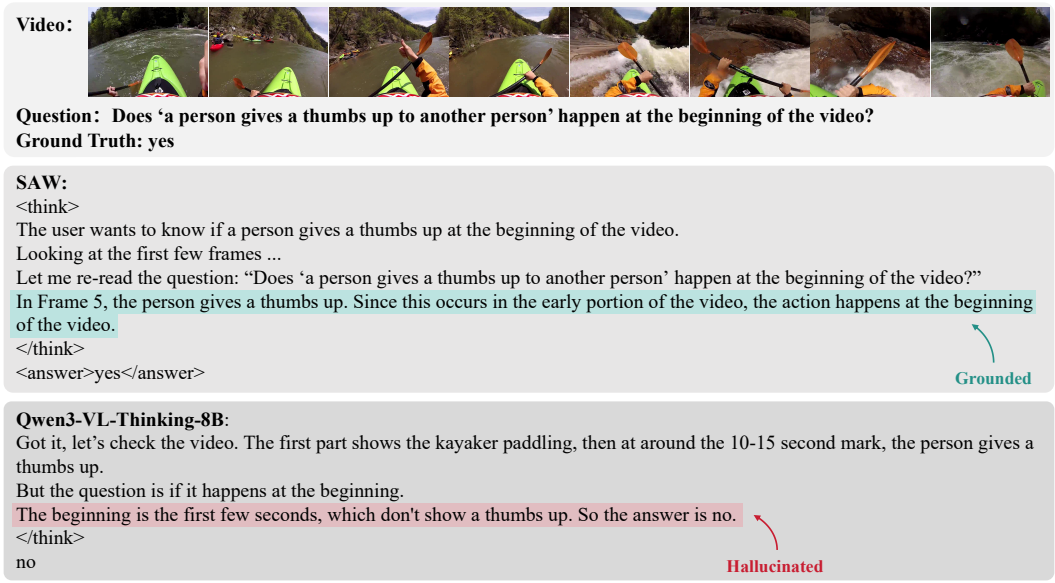

Figure 9: **Case 1: Event localization (kayaking).** The baseline hallucinates that the thumbs-up occurs later; SAW correctly identifies it in Frame 5.

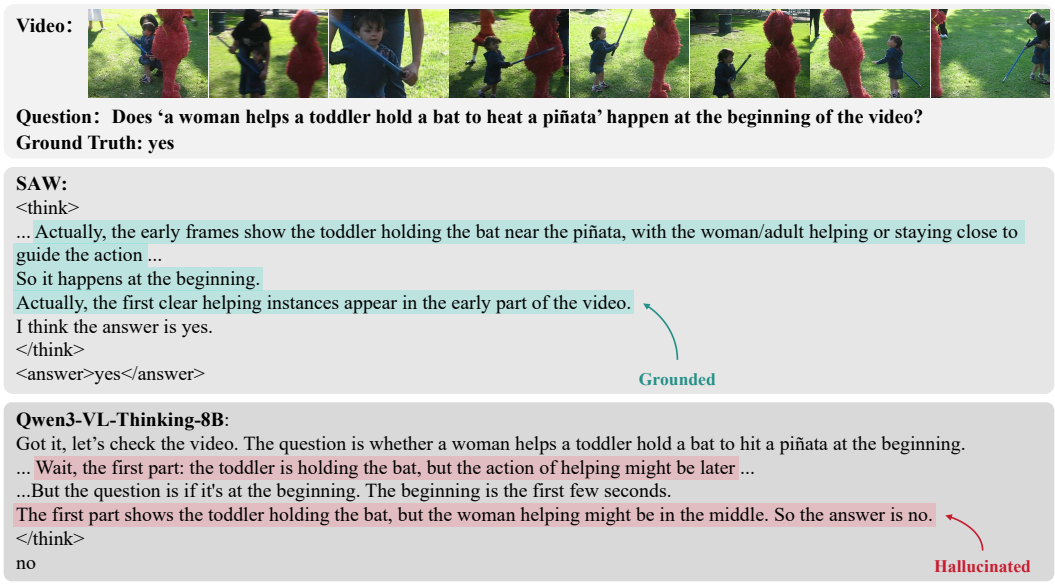

Figure 10: **Case 2: Event localization (piñata).** The baseline speculates the helping action occurs in the middle; SAW grounds it in the early frames.

Temporal order reasoning. Figures 11–12 ask whether one event occurs *later than* another. SAW's final "No" answers match the reference, but its explanations require caution. The official paired annotations place the credits before cutting in Case 3, and skiing before the costume event in Case 4. SAW instead claims that credits are absent or the events are simultaneous. These cases demonstrate answer agreement, not verified grounding of the reasoning traces.

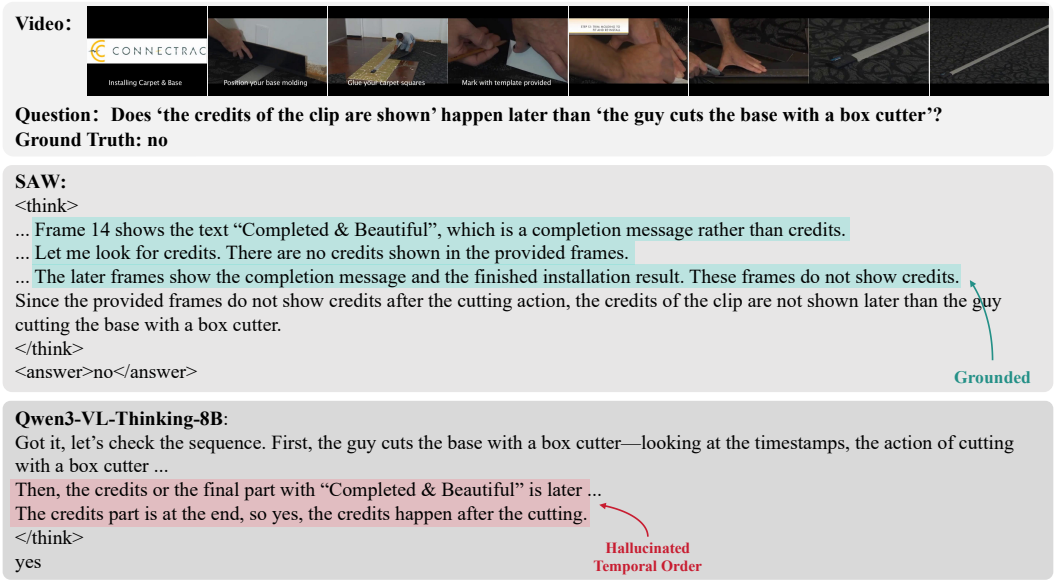

Figure 11: **Case 3: Temporal order (credits vs. cutting).** SAW answers "No," matching the reference; its claim that credits are absent is not supported by the dataset's paired annotation.

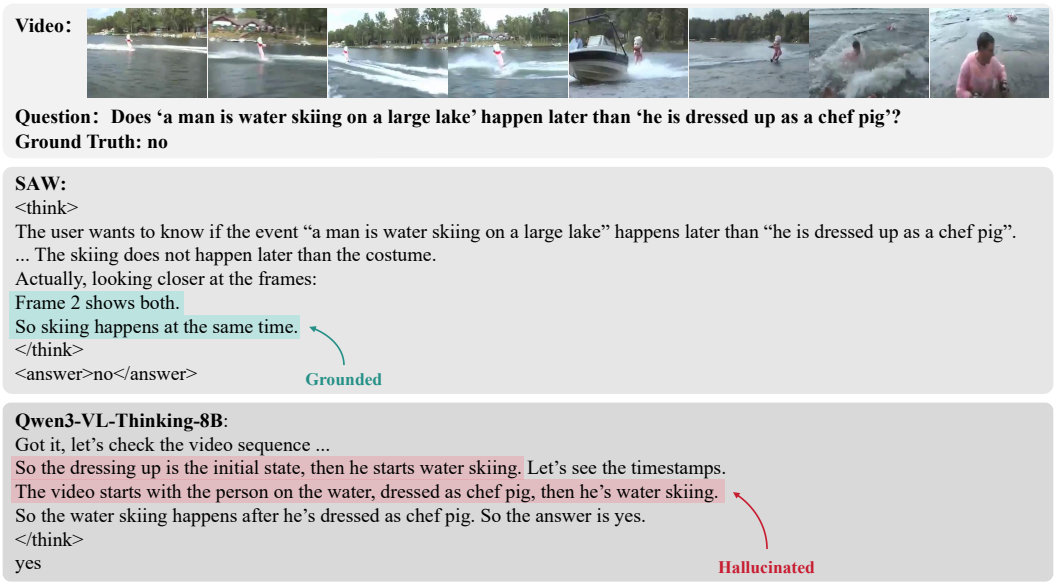

Figure 12: **Case 4: Temporal order (skiing vs. costume).** SAW answers "No," matching the reference; its simultaneity explanation differs from the dataset's earlier/later annotation.