CF-GRPO: COUNTERFACTUAL GROUP RELATIVE POLICY OPTIMIZATION FOR VIDEO HALLUCINATION SUPPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video language models have made substantial progress in video understanding, yet they remain prone to hallucination, producing answers that are inconsistent with the video content. Reinforcement fine-tuning with verifiable outcome rewards has been explored to mitigate such hallucinations. However, these rewards cannot distinguish visually grounded correct answers from those produced by the language prior. We show that this *source-blind credit assignment* is inherent in outcome-reward GRPO: the expected update is determined by the probability of correctness, regardless of whether that correctness depends on visual evidence. Our analysis further motivates interventions at three complementary levels of the GRPO update dynamics: the direction of pair-level updates, the magnitude of sample-level reinforcement, and the allocation of credit within each output group. Accordingly, we propose CF-GRPO, with three mechanisms targeting these respective levels. Specifically, 1) At the *pair level*, counterfactual pair optimization jointly trains opposite-label prompts sharing the same video, cancelling the common answer-shift update of prompts that the policy answers alike while preserving a positive discriminative update. 2) At the *sample level*, blind-failure advantage shaping scales the GRPO advantage by the model's failure probability without the video, reducing reinforcement for samples already solvable from the language prior. 3) At the *within-group level*, an evidence-consistency reward distinguishes outputs with identical correctness by whether their reasoning states answer-consistent evidence, restoring learning signal when outcome-reward GRPO provides none. Experiments on three video hallucination benchmarks show that CF-GRPO consistently improves over GRPO and strong baselines, while preserving the model's answer distribution and general video understanding.

1 INTRODUCTION

Recent video language models have made substantial progress in video understanding, with increasingly strong capabilities in event recognition, question answering, and reasoning over long videos Lin et al. (2024); Li et al. (2024b;a); Bai et al. (2025); Tang et al. (2025). Despite these advances, they remain prone to hallucination, producing answers that are inconsistent with the visual content of the input video Wang et al. (2024b); Zhang et al. (2024a). One important source of such errors is insufficient dependence on the video: the model may instead rely on its language prior, namely statistical associations between entities and events acquired during pretraining Zhang et al. (2024a). Crucially, however, prior-driven behavior is not always incorrect. Because these associations often reflect common event co-occurrences, the answer suggested by the prior may happen to agree with the video. As illustrated in Figure 1, a model can output the identical correct answer both with visual frames and without them (relying solely on text priors). Thus, answer correctness alone does not reveal whether a prediction is genuinely grounded in visual evidence.

To mitigate video hallucination, recent studies have explored reinforcement fine-tuning Pan et al. (2026), building on the broader adoption of group relative policy optimization (GRPO) for video language models Feng et al. (2026); Li et al. (2025c). GRPO samples multiple outputs for each prompt and reinforces them according to rewards determined by the correctness of their final answers. However, such outcome rewards observe only whether an answer is correct, not how that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: *Source-blind credit assignment* under outcome rewards. The model gives the same correct answer with and without the video. Outcome rewards assign identical credit to a visually grounded correct answer and one obtainable from the language prior alone.

answer was obtained. A correct answer grounded in visual evidence therefore receives the same credit as one that can already be produced from the language prior. We show that this ambiguity is inherent in standard GRPO with outcome rewards. In expectation, the update toward the correct answer depends only on the policy's probability of being correct; whether the same answer can already be produced without the video does not enter the update. Consequently, when prior-driven behavior happens to be correct, outcome-reward GRPO cannot distinguish it from visually grounded behavior and assigns it the same reinforcement. We refer to this failure as *source-blind credit assignment*.

To address this challenge, we propose CF-GRPO, which restructures the GRPO update dynamics to distinguish correctness that depends on visual evidence from correctness that remains obtainable without it. We demonstrates that this distinction can be enforced at three complementary levels: the direction of pair-level updates, the magnitude of sample-level reinforcement, and the allocation of credit within each output group. 1) At the *pair level*, counterfactual pair optimization jointly trains two prompts that share the same video but have opposite labels. We show that, when the policy assigns the pair the same answer distribution, their common answer-shift update cancels while the discriminative update remains positive. The pair therefore suppresses a direction that can be followed without distinguishing the two events, while preserving the direction that separates them using the video. 2) At the *sample level*, blind-failure advantage shaping estimates how often the policy fails on each prompt when the video is replaced by a blank input and uses this failure rate to scale its GRPO advantage. This attenuates reinforcement for samples that can already be answered correctly without visual input, while retaining larger updates where the video is needed. 3) At the *within-group level*, an evidence-consistency reward assigns credit according to both answer correctness and evidence mentions in the reasoning. Unlike an outcome reward, which provides no learning signal when all outputs have same correctness, this reward can distinguish outputs through the evidence consistency.

To validate CF-GRPO, we conduct experiments on three video hallucination benchmarks under a matched protocol, where every trained variant shares the same initialization, training videos, prompt budget, and learning rate. Built on Qwen2.5-VL-7B, CF-GRPO improves over the base model by 8.1% on average, with gains of up to 11.5% on a single benchmark, and outperforms outcome-reward GRPO trained under the same protocol by 2.6% on average, also surpassing the released OMNEX-VL checkpoint on every benchmark. Crucially, these gains are achieved without shifting the answer distribution of the model and without loss of general video understanding, so they reflect better discrimination rather than a preference for either answer.

2 Related Work

Multimodal large language models for video understanding. Early video language models process a fixed number of frames and aggregate their features before feeding them to the language model Maaz et al. (2024); Li et al. (2025b); Lin et al. (2024), which limits them to short clips. Later models

support longer videos by reducing visual tokens Cheng et al. (2024); Li et al. (2024a), extending the language-model context window Zhang et al. (2024b), or using variable-rate frame sampling with temporal position encoding Wang et al. (2024a); Bai et al. (2025). Across these models, the language model contains most parameters and is pretrained primarily on text, so language priors can influence predictions alongside visual evidence.

Reinforcement learning for video language models. Group relative policy optimization Shao et al. (2024) with verifiable outcome rewards has been extended from text reasoning to video. VideoChat-R1 rewards answer accuracy and output format for video question answering Li et al. (2025c), DeepVideo-R1 turns the GRPO objective into a regression on the advantage Park et al. (2026), and TW-GRPO Dang et al. (2026) weights the tokens of an output by their variation across the group. All three credit an output by its final answer, so a correct answer produced from the language prior is credited like one produced from the video. A second line of work makes the reward sensitive to the video by altering the video. Video-R1 adds a bonus when a group of outputs is more accurate with the frames in order than with the frames shuffled Feng et al. (2026), and CRPO flips or reverses the video and rewards answers that change accordingly Huang et al. (2026). Both supply samples on which the answer from the prior is wrong, but the outcome reward on the original video is unchanged, so the prior is still reinforced wherever it answers the original video correctly.

Hallucination in video language models. Recent benchmarks evaluate hallucinations involving objects, events, and temporal relations in video language models Wang et al. (2024b); Zhang et al. (2024a); Li et al. (2025a). DPO-based methods mitigate these errors through preference comparisons between responses or video inputs. VistaDPO incorporates spatial and temporal alignment into preference optimization Huang et al. (2025), while PaMi-VDPO constructs video preferences through prompt-aware augmentations during training Ding et al. (2025). For video question answering with verifiable answers, reinforcement learning allows the model to learn directly from the correctness of responses sampled during training. OMNEX-VL further introduces caption feedback, rewarding a generated caption when a separate language model answers correctly from the caption and question Pan et al. (2026). However, correctness rewards can also credit answers obtainable from the language prior alone. Our work addresses this limitation within GRPO by directing reinforcement to answers and reasoning that depend on the video.

3 METHOD

CF-GRPO is motivated by a credit-assignment view of video hallucination: outcome-reward GRPO reinforces correctness without distinguishing whether it depends on visual evidence or is already obtainable from the language prior. We first formalize this *source-blind credit assignment* in Section 4.1. Guided by this analysis, we restructure the GRPO update dynamics at three complementary levels, as illustrated in Figure 2. *1) At the pair level* (Section 3.2), counterfactual pair optimization controls the update direction by suppressing common answer-shift updates while preserving discriminative updates. *2) At the sample level* (Section 3.3), blind-failure advantage shaping controls the reinforcement magnitude by downweighting samples that the current policy can already solve without visual input. *3) At the within-group level* (Section 3.4), the evidence-consistency reward controls credit allocation by distinguishing outputs according to whether their reasoning states answer-consistent evidence. Together, these mechanisms make reinforcement sensitive to whether correctness depends on visual evidence, rather than to correctness alone.

3.1 PROBLEM SETUP AND SOURCE-BLIND CREDIT ASSIGNMENT

Group Relative Policy Optimization. GRPO Shao et al. (2024) trains a policy π_θ from a reference policy π_{ref}. For each prompt q sampled from the dataset distribution $P(Q)$, it generates a group of G outputs $\{o_1, \ldots, o_G\}$ from $\pi_{\theta_{\text{old}}}$, the policy before the current update, and maximizes

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q,\,\{o_i\}}\Big[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\big(\rho_{i,t}A_i,\ \text{clip}(\rho_{i,t}, 1\pm\epsilon)\,A_i\big) - \beta\, D_{\text{KL}}[\pi_\theta \,\|\, \pi_{\text{ref}}]\Big], \quad (1)$$

where $|o_i|$ is the length of o_i, $\rho_{i,t}$ is the probability of its t-th token under π_θ divided by that under $\pi_{\theta_{\text{old}}}$, ϵ is the clipping range, and β weights the KL divergence from π_θ to π_{ref}. The advantage of o_i

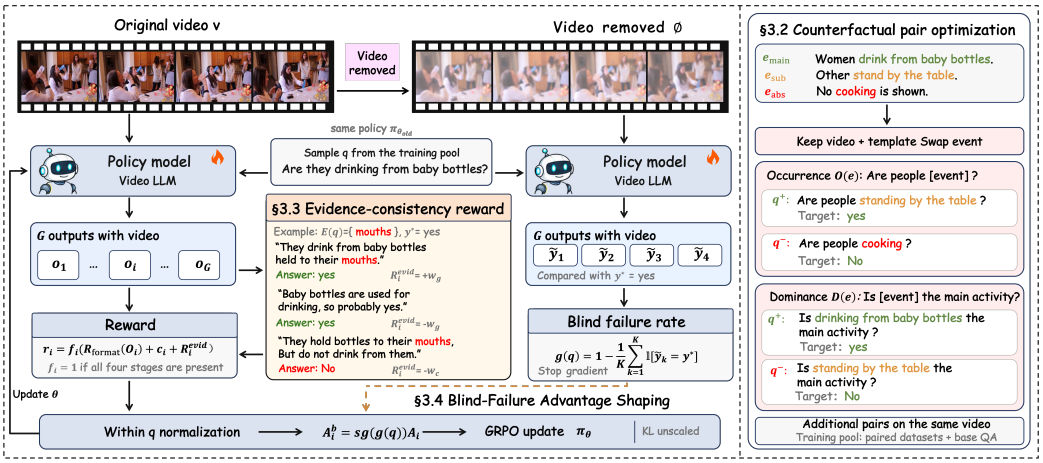


Figure 2: **Overview of CF-GRPO.** CF-GRPO addresses source-blind credit assignment at three complementary levels of the GRPO update dynamics. At the pair level, **counterfactual pair optimization** (§3.2) suppresses common answer-shift updates while preserving discriminative updates. At the sample level, **blind-failure advantage shaping** (§3.3) downweights samples already solvable without visual input. At the within-group level, the **evidence-consistency reward** (§3.4) reallocates credit according to answer-consistent evidence.

is its reward r_i standardized within the group,

$$A_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\})}. \tag{2}$$

Problem Setup. Following prior work on video hallucination Wang et al. (2024b); Zhang et al. (2024a); Li et al. (2025a), we formulate the task as binary video question answering. A prompt q asks whether an event occurs in a video v, and the policy generates $o = (z, y) \sim \pi_\theta(o \mid v, q)$, consisting of a chain of thought z and an answer $y \in \{\text{yes}, \text{no}\}$. The ground-truth answer y^* is yes if the queried event occurs in v and no otherwise, and we call (v, q, y^*) a training sample. In this setting, the policy hallucinates when it predicts yes although $y^* =$ no, a behavior that prior work associates with reliance on the language prior Zhang et al. (2024a). Following standard practice Shao et al. (2024), we use a rule-based outcome reward

$$r_i = R_{\text{format}}(o_i) + c_i, \qquad c_i = \mathbb{1}[y_i = y^*], \tag{3}$$

where R_{format} checks whether o_i follows the required format. The reward therefore observes only whether the final answer is correct, not the reasoning z_i or how the answer was obtained.

Source-Blind Credit Assignment. The language prior itself is not directly observable, so we operationalize prior-driven behavior by removing the visual input. Specifically, we replace v with a blank video $\varnothing$, a zero tensor of the same shape as the video input, and examine the answer distribution $\pi_\theta(y \mid \varnothing, q)$. We regard an answer as not depending on the video when

$$\pi_\theta(y \mid v, q) \approx \pi_\theta(y \mid \varnothing, q). \tag{4}$$

When such an answer is correct, the policy can obtain the outcome reward without relying on visual evidence. We next characterize how this ambiguity propagates from the reward into the GRPO update. Let $s(q)$ denote the probability that the policy answers prompt q correctly.

Proposition 1 (Source-blind outcome update)**.** *Under the analysis setting of Appendix A.1, the expected policy-gradient component of outcome-reward GRPO along logit of the correct answer is*

$$\sigma(q) = \sqrt{s(q)\big(1 - s(q)\big)}. \tag{5}$$

Hence, the expected update is determined solely by the correctness probability $s(q)$; whether the same correct answer depends on video or obtainable without the video does not enter the update.

The proof is given in Appendix A.1.

Remark 1. *Proposition 1 establishes that the ambiguity in the outcome reward persists in the policy update: when an answer obtainable without the video happens to be correct, it receives the same reinforcement as a visually grounded correct answer. We refer to this inability to condition reinforcement on the source of correctness as source-blind credit assignment.*

3.2 Counterfactual Pair Optimization

Proposition 1 shows that outcome-reward GRPO cannot distinguish visually grounded correctness from correctness obtainable from the language prior. To counter such prior-driven behavior, we train cases where the answer obtainable without the video is incorrect. However, training these cases independently can introduce a new shortcut: label imbalance may induce a common shift toward one answer without requiring the model to distinguish the video content. We therefore jointly train positive- and negative-label prompts and analyze the direction of their combined update.

Definition 2 (Pairwise update coordinates)**.** *Let q^+ and q^- be prompts with positive and negative labels that are trained together, and let $\ell(q)$ denote the logit of the positive answer for prompt q. We define the* common answer-shift coordinate *and the* discriminative coordinate *as*

$$a = \frac{\ell(q^+) + \ell(q^-)}{2}, \qquad d = \ell(q^+) - \ell(q^-). \tag{6}$$

The coordinate a moves the positive-answer logits of both prompts together, whereas d measures their separation.

Proposition 2 (Selective update under paired training)**.** *Under the outcome-reward update of Proposition 1, the expected gradient components along a and d are*

$$U_a = \sigma(q^+) - \sigma(q^-), \qquad U_d = \frac{1}{2}\left[\sigma(q^+) + \sigma(q^-)\right]. \tag{7}$$

If the policy assigns both prompts the same positive-answer probability $p \in (0, 1)$, then

$$U_a = 0, \qquad U_d = \sqrt{p(1-p)} > 0. \tag{8}$$

The proof is given in Appendix A.1.

Remark 2 (Selective cancellation)**.** *Proposition 2 demonstrates that paired training does not simply reduce the update. When the two prompts receive the same answer distribution, their common answer-shift component cancels while a positive discriminative component remains. Thus, if a pair is chosen so that the language prior treats its two prompts alike, joint training removes the update that follows their shared prior preference while preserving the update that separates them.*

Counterfactual Pair Construction. To approximate the condition in Proposition 2, we construct counterfactual pairs that share the same video and question template but have opposite ground-truth labels. For a training video v, let e_{main} denote its dominant event, e_{sub} a secondary event that occurs but is not dominant, and e_{abs} a plausible event that does not occur. Let $O(e)$ ask whether event e occurs in the video and $D(e)$ ask whether e is the dominant event. We construct

$$\mathcal{P}(v) = \left\{ \big(O(e_{\text{sub}}), O(e_{\text{abs}})\big),\ \big(D(e_{\text{main}}), D(e_{\text{sub}})\big) \right\}, \tag{9}$$

where the first prompt of each pair has a positive label and the second has a negative label. Within each pair, the two prompts share the same video and question template but query different, scene-plausible events with opposite labels, encouraging discrimination based on the video rather than prior preference. Moreover, e_{sub} is positive under O but negative under D, preventing event identity from serving as a fixed label shortcut. Appendix A.6 lists the prompts that build the pairs. The two members of each pair are always sampled and trained together.

3.3 Blind-Failure Advantage Shaping

Counterfactual pairing controls the direction of the joint update, but it does not determine how strongly each individual sample should contribute. The training pool still contains samples that the current policy can answer correctly without the video, and outcome-reward GRPO continues to reinforce them. Moreover, whether a sample is answerable without visual input depends on the current policy and changes throughout training, so it cannot be determined once in advance. We estimate this dependence at every training step and use it to scale the sample-level advantage.

Definition 3 (Blind failure weight)**.** *For a sample q, let*

$$b(q) = \Pr_{\pi_{\theta_{\text{old}}}}\big(y = y^* \mid \varnothing, q\big) \tag{10}$$

denote the probability that the current policy answers correctly without the video. We define its blind failure weight *as $1 - b(q)$. At each training step, we estimate this quantity using K blind outputs $\tilde{o}_k \sim \pi_{\theta_{\text{old}}}(\cdot \mid \varnothing, q)$:*

$$g(q) = 1 - \frac{1}{K}\sum_{k=1}^{K} \tilde{c}_k, \qquad \tilde{c}_k = \mathbb{1}[\tilde{y}_k = y^*], \tag{11}$$

so that $g(q)$ estimates $1 - b(q)$.

We use this weight to scale the GRPO advantage of every output generated with the video,

$$A_i^b = \text{sg}(g(q))\, A_i, \tag{12}$$

where sg denotes stop gradient. The blind outputs therefore affect training only through the scalar $g(q)$ and receive no gradient.

Proposition 3. *Under the analysis setting of Appendix A.1, replacing A_i with A_i^b scales the expected gradient component along the correct-answer logit of a sample q to $\big(1 - b(q)\big)\sigma(q)$ For a counterfactual pair (q^+, q^-), the expected gradient components along a and d is*

$$\widehat{U}_a = \big(1-b(q^+)\big)\sigma(q^+) - \big(1-b(q^-)\big)\sigma(q^-), \quad \widehat{U}_d = \tfrac{1}{2}\Big[\big(1-b(q^+)\big)\sigma(q^+) + \big(1-b(q^-)\big)\sigma(q^-)\Big]. \tag{13}$$

If the policy answers both samples from the prior with probability p of the positive answer, then $\widehat{U}_a = (1-2p)\sqrt{p(1-p)}$ and $\widehat{U}_d = \frac{1}{2}\sqrt{p(1-p)}$.

The proof is given in Appendix A.1.

Remark 3 (Adaptive magnitude control)**.** *Proposition 3 demonstrates that blind-failure shaping controls the magnitude of reinforcement at the sample level. For a sample with $b(q) = 1$, the policy-gradient contribution vanishes, whereas for $b(q) = 0$ the original outcome-reward update is retained; intermediate values interpolate between the two. Thus, samples already solvable without visual input receive less reinforcement, while samples the policy fails without the video retain larger updates.*

3.4 Evidence-Consistency Reward

Credit within a Group. The scaling of Section 3.3 acts on a sample as a whole and leaves untouched how credit is divided among the G outputs of the sample. This division is set by the reward, and the reward of Section 3.1 depends only on correctness. A correct output therefore receives the same reward whether or not its reasoning is based on the video, and a group in which every answer is correct or every answer is wrong receives no policy-gradient update regardless of how its outputs differ in their reasoning. Concretely, let $\boldsymbol{\phi}_i = |o_i|^{-1}\sum_t \nabla_\theta \log \pi_\theta(o_{i,t} \mid v, q, o_{i,<t})$ denote the length-normalized score of output i. At the sampling policy, the policy-gradient part of the update of one sample under Equation 1 is

$$\mathbf{u}_q = \frac{\text{Cov}(r, \boldsymbol{\phi})}{\text{std}(r)}, \qquad \text{Cov}(x, \boldsymbol{\phi}) = \frac{1}{G}\sum_{i=1}^{G}(x_i - \bar{x})\,\boldsymbol{\phi}_i. \tag{14}$$

We add a reward term that separates outputs by whether their reasoning states the evidence.

Evidence Keys and Reward. Each sample carries a set of evidence keys $E(q)$ identifying the video content that determines the answer. We set $e_i = 1$ if an output mentions a key in its reasoning before the answer, and 0 otherwise. Since counterfactual prompts may themselves contain these keys, rewarding key mentions alone would permit copying without visual grounding. We therefore reward evidence mentions only when the answer is correct and penalize them when the answer is wrong. Let f_i be a format gate that equals 1 for a valid output and 0 otherwise:

$$r_i = f_i\big(R_{\text{format}}(o_i) + c_i + R_i^{\text{evid}}\big), \qquad R_i^{\text{evid}} = w_g\, c_i\,(2e_i - 1) - w_c\,(1 - c_i)\,e_i, \tag{15}$$

with $w_g, w_c > 0$. Thus, a correct output gains w_g for stating the evidence and loses w_g otherwise, while a wrong output loses w_c for stating inconsistent evidence.

Proposition 4. *Let every output of a group pass the format check. Under the outcome reward,* $\mathrm{Cov}(r, \boldsymbol{\phi}) = \mathrm{Cov}(c, \boldsymbol{\phi})$*, so a group in which every answer is correct or every answer is wrong yields no policy-gradient update. Under the reward of Equation 15,*

$$\mathrm{Cov}(r, \boldsymbol{\phi}) = (1 - w_g)\,\mathrm{Cov}(c, \boldsymbol{\phi}) + (2w_g + w_c)\,\mathrm{Cov}(ce, \boldsymbol{\phi}) - w_c\,\mathrm{Cov}(e, \boldsymbol{\phi}), \tag{16}$$

and with the scaling of Section 3.3 the update of a group in which every answer is correct is $\widehat{\mathbf{u}}_q = g(q)\,\mathrm{Cov}(e, \boldsymbol{\phi})/\mathrm{std}(e)$*, while that of a group in which every answer is wrong is* $-g(q)\,\mathrm{Cov}(e, \boldsymbol{\phi})/\mathrm{std}(e)$.

Proposition 4 is proved in Appendix A.1.4.

Remark 4. *Proposition 4 confirms that the evidence-consistency reward restores within-group learning signals even when correctness is uniform: it promotes evidence-consistent reasoning for correct answers and suppresses it for incorrect ones, while preserving sample-level scaling by* $g(q)$.

4 Experiments

4.1 Experimental Settings

Implementation details. We use Qwen2.5-VL-7B-Instruct Bai et al. (2025) as the base model and train for 300 steps on eight A800 GPUs with $G = 8$ outputs and $K = 4$ blind outputs per sample, a learning rate of 3×10^{-6}, $\beta = 0.1$, $w_g = 0.4$, and $w_c = 0.3$. Videos are sampled at 2 frames per second with a per-frame budget of $401{,}408$ pixels, following OMNEX-VL Pan et al. (2026), and evaluation uses 32 frames with greedy decoding.

Training data. The videos come from LLaVA-Video-178K and share no file with the benchmarks. The pool holds $3{,}058$ base samples, whose questions with answer no are made by changing one attribute of a caption sentence, and $2{,}000$ samples in counterfactual pairs from 500 videos, balanced between the two labels. Appendix A.4 describes how the base samples and the pairs are built, and Appendix A.6 lists the prompts.

Baselines. SFT denotes the supervised model from which every reinforcement fine-tuning run starts. GRPO uses the outcome reward of Section 3.1 on a generic balanced set of samples from the same captions. OMNEX-VL Pan et al. (2026) is the released checkpoint. All models are evaluated by us under the same frames and decoding, the three open-source video LLMs with their official inference code and default settings.

Evaluation benchmarks. We evaluate on three video hallucination benchmarks. VideoHallucer Wang et al. (2024b) scores a question pair only when both members are answered correctly, EventHallusion Zhang et al. (2024a) asks binary event questions over entire, mixed, and misleading videos, and VidHalluc Li et al. (2025a) probes action confusions between similar clips with its official score. General capability is checked on MMVU Zhao et al. (2025), Video-MMMU Hu et al. (2026), and MMBench Liu et al. (2024).

4.2 Comparison with State-of-the-Art Methods

Hallucination benchmarks. Table 1 compares CF-GRPO with open-source video LLMs and with models trained from the same SFT model on VideoHallucer Wang et al. (2024b), EventHallusion Zhang et al. (2024a), and VidHalluc Li et al. (2025a), whose metrics are described in Appendix A.2. We observe three findings. *(i) CF-GRPO has the best overall score of every benchmark among the trained models*, ahead of GRPO with the outcome reward by 3.91, 3.91, and 0.12 and of OMNEX-VL by 5.38, 1.52, and 4.09. *(ii) The gain does not come from a shift toward either answer.* On VideoHallucer, CF-GRPO is equally accurate on positive and negative questions, 76.04 against 75.26. *(iii) A single overall number can reward a prior.* The base model owes its EventHallusion score to the entire and misleading subsets, whose answers are all no, and fails the mixed subset. VideoChat2 has a prior toward yes and reaches the highest overall EventHallusion score with the lowest paired accuracy on VideoHallucer. CF-GRPO is the best trained model on the two subsets whose answers are all no and, by category of VideoHallucer (Appendix Table 5), leads on the two categories whose question with answer no describes a plausible but absent relation or event.

General benchmarks. Table 2 evaluates general capability on MMVU Zhao et al. (2025), Video-MMMU Hu et al. (2026), and MMBench Liu et al. (2024). CF-GRPO is the best trained model

Table 1: Comparison with state-of-the-art MLLMs on hallucination benchmarks.

	EventHallusion				VideoHallucer				VidHalluc			
Method	Entire	Mix	Misl.	Overall	Obj.	Sem.	Ext.	Overall	Acc	Yes	No	ACH
Open source video LLMs												
Video-LLaVA-7B	35.09	27.46	75.49	41.56	52.50	30.00	29.50	25.73	73.02	70.86	75.19	48.44
VideoChat2-7B	36.84	70.98	85.29	65.04	39.00	32.00	15.50	18.35	70.23	98.97	41.50	40.91
LLaVA-OneVision-7B	58.77	47.67	92.16	61.86	55.00	61.50	50.00	46.25	88.32	92.75	83.89	77.22
Qwen2.5-VL-7B	64.04	37.31	95.10	59.17	43.00	53.00	43.50	46.48	82.68	73.04	92.33	67.42
Training on Qwen2.5-VL-7B												
SFT	47.37	49.22	75.49	55.26	63.50	64.50	57.00	52.60	86.06	89.52	82.60	73.50
GRPO	40.35	56.48	81.37	58.19	62.00	**73.50**	53.50	52.44	89.05	93.90	84.20	78.78
OMNEX-VL	41.23	53.37	80.39	56.72	60.50	68.00	56.50	54.83	86.92	93.67	80.16	74.81
CF-GRPO (ours)	**48.25**	**59.59**	**82.35**	**62.10**	**64.50**	73.00	**58.75**	**56.35**	**89.17**	92.80	**85.53**	**78.90**

Table 2: Comparison with the state-of-the-art MLLMs on general benchmarks.

	MMVU	**Video-MMMU**				**MMBench (en)**		
Method	Overall↑	Perc.↑	Comp.↑	Adapt.↑	Overall↑	Action↑	Spatial↑	Overall↑
Qwen2.5-VL-7B	55.36	61.00	44.67	43.01	49.72	93.0	72.9	87.48
SFT	63.84	67.67	43.67	46.95	52.90	94.0	67.8	88.08
GRPO	63.68	69.00	**44.00**	46.59	53.36	94.4	67.2	88.15
OMNEX-VL	64.32	66.00	**44.00**	43.37	51.31	94.0	65.5	87.96
CF-GRPO (ours)	**65.44**	**70.67**	43.67	**47.31**	**54.04**	**94.9**	**68.4**	**88.03**
△ vs. GRPO	(+1.8)	(+1.7)	(−0.3)	(+0.7)	(+0.7)	(+0.5)	(+1.2)	(−0.1)

on MMVU and Video-MMMU and within 0.2 of the best on MMBench, so it gives up no general capability relative to the other trained models.

4.3 Ablation Study

Component analysis. Table 3 ablates the components of CF-GRPO cumulatively. Removing the counterfactual pairs lowers all three benchmarks, EventHallusion by 5.38, because the pool then contains few samples on which the answer from the prior is wrong. The two terms of the evidence-consistency reward act on different benchmarks. Setting $w_c = 0$ lowers VideoHallucer by a further 0.79 and raises the share of outputs that state the evidence yet answer against it from 3.63 to 4.35, whereas also setting $w_g = 0$ lowers EventHallusion by a further 1.96. Removing the shaping next raises EventHallusion by 5.38 but lowers VideoHallucer by 2.72, since the entire and misleading subsets of EventHallusion reward a drift toward the negative answer while the paired metric of VideoHallucer penalizes it, so the shaping keeps the answer distribution in place. Replacing the original sample of each pair by a paraphrase of the dominant event lowers EventHallusion to 56.23 and VideoHallucer to 53.90, because rewarding a restated event teaches the model to accept restatements, which the counterfactual samples cannot counteract. The blind outputs raise the cost of a step by half, but at equal rollout budgets CF-GRPO still leads GRPO, 62.10 against 59.17 on EventHallusion and 56.35 against 54.21 on VideoHallucer.

Answer distribution. A gain in accuracy is evidence of better discrimination only if the answer distribution does not move. Figure 3 places every model of Table 1 in the plane of accuracy on positive prompts against accuracy on negative prompts. SFT lies above the diagonal, GRPO and OMNEX-VL lie below it by 7.6 and 7.8, and CF-GRPO ends within one of it with the highest paired accuracy. On VidHalluc, CF-GRPO raises both accuracies of Table 1 over SFT and keeps their gap within 0.4 of SFT, whereas GRPO and OMNEX-VL widen it by 2.8 and 6.6. The shaping multiplies the update of each sample by a non-negative factor, and on a pair that the prior answers alike it adds a shift only against the answer the prior favors (Proposition 3), so the objective pulls the answer distribution toward balance rather than toward either answer.

Dependence on the video. Equation 4 concerns the answer distribution. As a decision-level proxy, we report in Table 4 the share of errors whose greedy answer does not change when the video is replaced by a zero tensor. For the base model, this share is 61.7, 78.2, and 76.4 percent on the three benchmarks, since without the video it gives the negative answer to nearly every question.

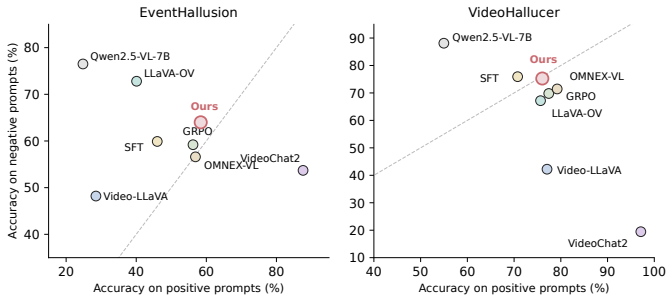


Figure 3: Accuracy on positive against negative prompts on EventHallusion and VideoHallucer.

Table 3: Ablation of CF-GRPO. Each row removes one component, and the small numbers give the change from the full model.

Method	EventHallusion	VideoHallucer	VidHalluc
CF-GRPO	62.10	56.35	78.90
− counterfactual pairs	56.72−5.38	54.36−1.99	76.44−2.46
− evidence penalty ($w_c = 0$)	60.15−1.95	53.57−2.78	77.52−1.38
− evidence reward ($w_g = w_c = 0$)	58.19−3.91	55.42−0.93	77.08−1.82
− advantage shaping	63.57+1.47	52.70−3.65	79.70+0.80
− our base samples	58.19−3.91	52.44−3.91	78.78−0.12

Table 4: Dependence of answers on the video, measured by replacing the video with zeros.

	EventHallusion			VideoHallucer			VidHalluc		
Method	Video	Unch.	Prior	Video	Unch.	Prior	Video	Unch.	Prior
Qwen2.5-VL-7B	59.2	61.7	25.2	71.1	78.2	22.6	82.8	76.4	13.2
SFT	55.3	48.6	21.8	72.6	57.6	15.8	86.0	46.0	6.4
GRPO	58.2	50.9	21.3	72.9	59.6	16.1	89.0	34.4	3.8
OMNEX-VL	56.7	45.2	19.6	74.6	50.0	12.7	86.9	43.4	5.7
CF-GRPO (ours)	62.1	42.6	16.1	75.0	52.7	13.2	89.1	35.5	3.9

Training lowers this share, and CF-GRPO lowers it more than GRPO, to 42.6 against 50.9 percent on EventHallusion and 52.7 against 59.6 percent on VideoHallucer, while the two are close on VidHalluc at 35.5 and 34.4 percent. Over all questions, the errors whose answer does not change without the video fall from 21.3 percent for GRPO to 16.1 percent for CF-GRPO on EventHallusion, and from 16.1 to 13.2 percent on VideoHallucer, close to the 12.7 percent of OMNEX-VL. The errors whose answer changes without the video differ by less than four points between the trained models. The gain over GRPO in Table 1 therefore comes from fewer errors whose answer does not change without the video.

5 Conclusion

Outcome-reward GRPO performs source-blind credit assignment, giving a correct answer the same reinforcement whether it comes from the video or from the language prior. CF-GRPO restructures the GRPO update at three levels to favor answers grounded in visual evidence. Counterfactual pair optimization jointly trains opposite-label prompts on the same video, cancelling their common answer-shift update while preserving the discriminative one. Blind-failure advantage shaping down-weights samples that the model answers correctly without the video, reducing reinforcement of the prior. The evidence-consistency reward distinguishes outputs with identical correctness by whether their reasoning states answer-consistent evidence, restoring a learning signal where outcome-reward GRPO provides none. CF-GRPO outperforms the baselines on three video hallucination benchmarks without shifting the answer distribution or degrading general video understanding.

AI USE STATEMENT

We used Claude and Chatgpt for writing assistance, including grammar correction, wording refinement, and minor stylistic editing of draft text. The model was not used for research ideation, data collection, dataset annotation, code generation, experimental design, or analysis. All technical content was authored and verified by the human authors, who take full responsibility for the content of this paper.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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

Jisheng Dang, Jingze Wu, Teng Wang, Xuanhui Lin, Nannan Zhu, Hongbo Chen, Wei-Shi Zheng, Meng Wang, and Tat-Seng Chua. Reinforcing video reasoning with focused thinking. In *European Conference on Computer Vision*, pp. 623–640. Springer, 2026.

Xinpeng Ding, Kui Zhang, Jianhua Han, Lanqing Hong, Hang Xu, and Xiaomeng Li. Pami-vdpo: Mitigating video hallucinations by prompt-aware multi-instance video preference learning. *arXiv preprint arXiv:2504.05810*, 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. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Xiang Yue, Bo Li, Yuanhan Zhang, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multidisciplinary professional videos. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27798–27828, 2026.

Haojian Huang, Haodong Chen, Shengqiong Wu, Meng Luo, Jinlan Fu, Xinya Du, Hanwang Zhang, and Hao Fei. Vistadpo: Video hierarchical spatial-temporal direct preference optimization for large video models. *arXiv preprint arXiv:2504.13122*, 2025.

Zhe Huang, Hao Wen, Aiming Hao, Bingze Song, Meiqi Wu, Jiahong Wu, Xiangxiang Chu, Sheng Lu, and Haoqian Wang. Taming hallucinations: Boosting mllms' video understanding via counterfactual video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8153–8163, 2026.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13723–13733. IEEE, 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, 2024b.

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.

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.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad 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.

Miao Pan, Wangjie Gan, Jintao Chen, Wenqi Zhang, Sun Bing, Jianwei Yin, and Xuhong Zhang. Ground what you see: Hallucination-resistant mllms via caption feedback, diversity-aware sampling, and conflict regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8242–8250, 2026.

Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo. *Advances in Neural Information Processing Systems*, 38:138605–138632, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2):1355–1376, 2025.

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.

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.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024b.

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

A Appendix

A.1 Proofs

Proposition 1 characterizes the expected update produced by standard outcome reward GRPO at the sampling policy. Appendix A.1.1 fixes the analysis setting, writes this update in the covariance form of Equation 14, and derives Equation 5. Propositions 2, 3, and 4 then characterize how the three changes introduced by CF-GRPO modify this update. Appendices A.1.2, A.1.3, and A.1.4 prove these propositions by evaluating the same covariance form under the corresponding changes.

A.1.1 Setting and Proof of Proposition 1

Setting. The training pool $\mathcal{D}$ consists of the base set and the counterfactual pairs. At the sampling policy, the CF-GRPO the policy-gradient component of the CF-GRPO update is that of Equation 1 with the scaled advantage and the evidence reward in place,

$$\widehat{\mathbf{g}}_{\mathrm{CF}}(\theta) = \mathbb{E}_{q\sim\mathcal{D}}\, \mathrm{sg}\big(g(q)\big)\, \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\frac{r_i-\bar{r}}{\mathrm{std}(r)}\,\nabla_\theta \log \pi_\theta(o_{i,t}\mid v, q, o_{i,<t}), \tag{17}$$

Apart from the two highlighted terms and the counterfactual pairs in $\mathcal{D}$, this component is that of GRPO. The KL term of Equation 1 is added unscaled, the blind outputs enter only through $g(q)$.

All statements concern the expected gradient of Equation 1 at the sampling policy $\pi_{\theta_{\mathrm{old}}}$. At this policy every ratio $\rho_{i,t}$ equals one and the clipping is inactive, so the gradient of $\min\big(\rho_{i,t}A_i, \mathrm{clip}(\rho_{i,t}, 1\pm\epsilon)A_i\big)$ is $A_i\nabla_\theta \log \pi_\theta(o_{i,t}\mid v, q, o_{i,<t})$ and the gradient of Equation 1 is

$$\nabla_\theta \mathcal{J}_{\mathrm{GRPO}} = \mathbb{E}_{q,\{o_i\}}\,\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} A_i\,\nabla_\theta \log \pi_\theta(o_{i,t}\mid v, q, o_{i,<t}) - \beta\,\nabla_\theta D_{\mathrm{KL}}[\pi_\theta \,\|\, \pi_{\mathrm{ref}}]. \tag{18}$$

Equation 17 is the first term of this gradient after the three changes: q is drawn from a pool that contains the pairs of Equation 9, r_i is computed by Equation 15, and A_i is multiplied by $\mathrm{sg}(g(q))$. The KL term is kept as in Equation 1. The three changes act only on the first term, so the statements below concern this term for one sample. With the advantage of Equation 2 and the length-normalized score $\boldsymbol{\phi}_i$ of Equation 14, the first term for one sample is

$$\mathbf{u}_q = \frac{1}{G}\sum_{i=1}^{G} A_i\boldsymbol{\phi}_i = \frac{1}{G}\sum_{i=1}^{G}\frac{r_i-\bar{r}}{\mathrm{std}(r)}\,\boldsymbol{\phi}_i = \frac{\mathrm{Cov}(r,\boldsymbol{\phi})}{\mathrm{std}(r)}, \tag{19}$$

which is Equation 14 with Cov as defined there. Two properties of this form are used in every proof: the covariance is linear in the reward, and it vanishes when the reward is constant within the group. We take the large-group limit, in which the mean and the standard deviation of any quantity over the group equal its mean and standard deviation under the policy. Outputs are taken to have the same length L, so the factor $1/|o_i|$ in $\boldsymbol{\phi}_i$ is the constant $1/L$, which we drop from the expected updates below. A group with constant reward has zero advantage by Equation 2, and every ratio $\mathrm{Cov}(\cdot,\boldsymbol{\phi})/\mathrm{std}(\cdot)$ below is set to zero when its denominator is zero. The implementation adds 10^{-4} to the standard deviation, which gives the same zero update for such groups. For a group in which every answer is correct or every answer is wrong, the large-group limit is taken under the policy conditioned on that correctness.

For Propositions 1, 2, and 3, the answer is the choice between the tokens yes and no at one position of the output, and the chain of thought before this position is held fixed, so $s(q)$ and $b(q)$ are the probabilities of the correct answer at this position with and without the video. Let $\ell(q)$ denote the logit of yes minus the logit of no at this position, so that $\pi_\theta(\text{yes}\mid q) = 1/(1+e^{-\ell(q)})$. This is the logit of the positive answer in Proposition 2. Let $\ell^*(q)$ be the logit of the correct answer, $\ell^*(q) = \ell(q)$ if $y^* = $ yes and $\ell^*(q) = -\ell(q)$ if $y^* = $ no. We treat each $\ell(q)$ as a free parameter. The component of $\mathbf{u}_q$ along $\ell^*(q)$ is then the partial derivative of the objective with respect to $\ell^*(q)$, and a gradient step of size η in these coordinates changes $\ell^*(q)$ by η times this derivative. In a network with shared parameters, the actual change is further scaled by $\|\nabla_\theta \ell^*(q)\|^2$ and interacts

across prompts, so the results below describe the gradient in the logit coordinates rather than the exact parameter update. Proposition 4 does not hold the chain of thought fixed, since e_i is read from it.

Proof of Proposition 1. Under the outcome reward, $r_i = R_{\text{format}}(o_i) + c_i$. When every output of the group passes the format check, R_{format} is constant within the group, so $\text{Cov}(r, \boldsymbol{\phi}) = \text{Cov}(c, \boldsymbol{\phi})$ and $\text{std}(r) = \text{std}(c) = \sqrt{s(1-s)}$, which is $\sigma(q)$ of Equation 5. The component of $\boldsymbol{\phi}_i$ along ℓ^* comes from the answer token, with $\partial \log \pi_\theta(y^*)/\partial \ell^* = 1 - s$ for a correct output and $\partial \log \pi_\theta(\neg y^*)/\partial \ell^* = -s$ for a wrong one. A fraction s of the group is correct with advantage $(1-s)/\sigma$ and a fraction $1-s$ is wrong with advantage $-s/\sigma$, so the component of $\mathbf{u}_q$ along ℓ^* is

$$s \cdot \frac{1-s}{\sigma}\,(1-s) \;+\; (1-s) \cdot \frac{-s}{\sigma}\,(-s) \;=\; \frac{s(1-s)}{\sigma} \;=\; \sigma(q). \tag{20}$$

A gradient step of size η in the logit coordinates therefore raises $\ell^*(q)$ by $\eta\,\sigma(q)$ to first order, which is Equation 5. Only $s(q)$ enters, so the update is the same whether the correct answers were produced from the video or from the prior. This is the reinforcement of the prior described in Section 3.2. Each proof below evaluates Equation 19 under one of the three changes.

A.1.2 Proof of Proposition 2

Proposition 2 evaluates Equation 20 on the two samples of a counterfactual pair. Each sample forms its own group, so pairing changes the update of neither sample. What pairing changes is the direction in which the two updates together move the two logits $\ell(q^+)$ and $\ell(q^-)$. Since $\ell^*(q^+) = \ell(q^+)$ and $\ell^*(q^-) = -\ell(q^-)$, Equation 20 gives the expected derivative $\sigma(q^+)$ along $\ell(q^+)$ and $-\sigma(q^-)$ along $\ell(q^-)$. In the coordinates of the proposition, $\ell(q^+) = a + d/2$ and $\ell(q^-) = a - d/2$, and the chain rule gives

$$U_a = \sigma(q^+) - \sigma(q^-), \qquad U_d = \tfrac{1}{2}\big(\sigma(q^+) + \sigma(q^-)\big), \tag{21}$$

which is Equation 7. With free logits, a gradient step changes a by $\eta U_a/2$ and d by $2\eta U_d$, so $U_a = 0$ and $U_d > 0$ also describe the change of the coordinates themselves. If the policy gives both samples the same answer distribution with probability p of the positive answer, then $s(q^+) = p$ and $s(q^-) = 1 - p$. Both samples then have $\sigma = \sqrt{p(1-p)}$, so $U_a = 0$ and $U_d = \sqrt{p(1-p)} > 0$ for $0 < p < 1$, which proves the proposition. A negative sample trained without its positive partner contributes $U_a = -\sigma(q^-)$ and $U_d = \frac{1}{2}\sigma(q^-)$ instead, so its update has a component along a, which the policy can follow without the video. Pairing removes this component whenever the two samples share an answer distribution, but it leaves the size $\sigma(q)$ of each update as it is, so a sample that the prior answers correctly still receives its full update. Appendix A.1.3 addresses the size.

A.1.3 Proof of Proposition 3

Proposition 3 evaluates Equation 20 under the advantages $\widehat{A}_i = \text{sg}(g(q))\,A_i$ of Equation 12. The blind outputs that define $g(q)$ are drawn independently of the outputs with the video, so the expectation of the scaled update factorizes, $\mathbb{E}[g(q)A_i\boldsymbol{\phi}_i] = \mathbb{E}[g(q)]\,\mathbb{E}[A_i\boldsymbol{\phi}_i]$, and $\mathbb{E}[g(q)] = 1 - \frac{1}{K}\sum_k \mathbb{E}[\tilde{c}_k] = 1 - b(q)$. The expected update along $\ell^*(q)$ is therefore $\big(1 - b(q)\big)\sigma(q)$, the first statement of the proposition. The chain rule of Appendix A.1.2 is linear in the two derivatives, so replacing $\sigma(q^\pm)$ by $\big(1 - b(q^\pm)\big)\sigma(q^\pm)$ in Equation 21 gives

$$\widehat{U}_a = \big(1-b(q^+)\big)\sigma(q^+) - \big(1-b(q^-)\big)\sigma(q^-), \quad \widehat{U}_d = \tfrac{1}{2}\Big[\big(1-b(q^+)\big)\sigma(q^+) + \big(1-b(q^-)\big)\sigma(q^-)\Big], \tag{22}$$

which is Equation 13. If the policy answers both samples from the prior with probability p of the positive answer, then by Equation 4 the answer distribution is the same with and without the video, so $b(q^+) = s(q^+) = p$ and $b(q^-) = s(q^-) = 1 - p$. With $\sigma = \sqrt{p(1-p)}$,

$$\widehat{U}_a = (1-p)\,\sigma - p\,\sigma = (1-2p)\,\sigma, \qquad \widehat{U}_d = \tfrac{1}{2}\sigma > 0, \tag{23}$$

which proves the proposition. $\widehat{U}_a$ is zero at $p = \frac{1}{2}$, negative when $p > \frac{1}{2}$, and positive when $p < \frac{1}{2}$. The shift along a is therefore directed against the answer from the prior, while $\widehat{U}_d > 0$ keeps the two samples separated. Pairing and shaping thus act on the update of a sample as a whole, the first on its direction and the second on its size. Neither changes how the update is divided among the G outputs of a sample. This division is what the evidence-consistency reward changes, and Appendix A.1.4 evaluates it.

A.1.4 Proof of Proposition 4

Proposition 4 concerns the division of the update among the outputs of one group, so the proof evaluates the full vector of Equation 19 rather than its component along $\ell^*(q)$. Let every output pass the format check, so that R_{format} is constant within the group. Under the outcome reward $r_i = R_{\text{format}} + c_i$, and linearity gives $\mathrm{Cov}(r, \boldsymbol{\phi}) = \mathrm{Cov}(c, \boldsymbol{\phi})$. In a group in which every answer is correct or every answer is wrong, c_i is constant, so every reward equals the group mean and the group yields no update however its outputs differ in their reasoning. Under the reward of Equation 15,

$$r_i = R_{\text{format}} + c_i + w_g\, c_i\,(2e_i - 1) - w_c\,(1 - c_i)\,e_i = R_{\text{format}} + (1 - w_g)\,c_i + (2w_g + w_c)\,c_i e_i - w_c\,e_i, \tag{24}$$

and linearity gives Equation 16. If every answer is correct, $c_i = 1$ and $r_i = R_{\text{format}} + 1 - w_g + 2w_g\,e_i$, so $\mathrm{Cov}(r, \boldsymbol{\phi}) = 2w_g\,\mathrm{Cov}(e, \boldsymbol{\phi})$ and $\mathrm{std}(r) = 2w_g\,\mathrm{std}(e)$. If every answer is wrong, $c_i = 0$ and $r_i = R_{\text{format}} - w_c\,e_i$, so $\mathrm{Cov}(r, \boldsymbol{\phi}) = -w_c\,\mathrm{Cov}(e, \boldsymbol{\phi})$ and $\mathrm{std}(r) = w_c\,\mathrm{std}(e)$. The scaling of Section 3.3 multiplies every advantage of the group by $g(q)$, so the two updates are $g(q)\,\mathrm{Cov}(e, \boldsymbol{\phi})/\mathrm{std}(e)$ and $-g(q)\,\mathrm{Cov}(e, \boldsymbol{\phi})/\mathrm{std}(e)$, which proves the proposition. If e_i is the same for every output, both vanish. The direction of the two updates follows from the identity $\mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(o)] = 0$. In a group in which every answer is correct, the outputs are drawn from $\pi_\theta(o \mid v, q, c = 1)$, and with the factor $1/L$ dropped, $\boldsymbol{\phi}_i = \nabla_\theta \log \pi_\theta(o_i \mid v, q)$. Under this distribution $\mathbb{E}[\boldsymbol{\phi}] = \nabla_\theta \log s(q)$ and $\mathbb{E}[e\,\boldsymbol{\phi}] = \nabla_\theta P_\theta(e = 1, c = 1 \mid v, q)/s(q)$, so in the large-group limit

$$\mathrm{Cov}(e, \boldsymbol{\phi}) = \mathbb{E}[e\,\boldsymbol{\phi}] - \mathbb{E}[e]\,\mathbb{E}[\boldsymbol{\phi}] = \nabla_\theta P_\theta(e = 1 \mid v, q, c = 1), \tag{25}$$

the gradient of the probability that a correct output states the evidence keys. The same identity with $c = 0$ holds for a group in which every answer is wrong. A group in which every answer is correct therefore raises the probability that a correct output states the evidence keys, scaled by $g(q)/\mathrm{std}(e)$, and a group in which every answer is wrong lowers the probability that a wrong output states them, with the same scale. The factor $g(q)$ carries through unchanged, so the three changes act on different parts of the update of a sample: pairing on its direction, shaping on its size, and the evidence term on its division among the outputs.

Table 5: VideoHallucer paired accuracy by category.

Method	Paired	Obj. Rel.	Temporal	Sem. Detail	Ext. Fact.	Ext. Nonfact.
Video-LLaVA-7B	25.73	52.50	1.14	30.00	15.50	29.50
VideoChat2-7B	18.35	39.00	2.27	32.00	3.00	15.50
LLaVA-OneVision-7B	46.25	55.00	39.77	61.50	25.00	50.00
Qwen2.5-VL-7B	46.48	43.00	60.23	53.00	32.66	43.50
SFT	52.60	63.50	50.00	64.50	28.00	57.00
GRPO	52.44	62.00	51.70	**73.50**	21.50	53.50
OMNEX-VL	54.83	60.50	**59.66**	68.00	**29.50**	56.50
CF-GRPO (ours)	**56.35**	**64.50**	57.95	73.00	27.50	**58.75**

Table 6: Accuracy with 8, 16, and 32 evaluation frames under the metrics of Table 1. Bold marks the best trained model in each column.

	EventHallusion			VideoHallucer			VidHalluc		
Frames	8	16	32	8	16	32	8	16	32
Qwen2.5-VL-7B	57.93	58.29	59.17	43.96	46.57	46.48	67.58	67.18	67.42
SFT	55.35	55.02	55.26	51.35	52.12	52.60	73.38	73.33	73.50
GRPO	58.02	58.32	58.19	51.95	52.08	52.44	78.32	78.41	78.78
OMNEX-VL	55.91	56.88	56.72	54.21	55.09	54.83	74.62	74.74	74.81
CF-GRPO (ours)	61.35	61.07	**62.10**	56.21	56.07	**56.35**	**78.69**	**78.48**	**78.90**

A.2 Evaluation Benchmarks and Metrics

VideoHallucer Wang et al. (2024b) pairs each positive question with a negative one on the same video and scores a pair only when both members are answered correctly. Table 1 reports this paired

accuracy overall and for three of its five categories, and Table 5 gives all five. EventHallusion Zhang et al. (2024a) asks binary event questions over entire, mixed, and misleading videos, and we report accuracy on each subset and overall. VidHalluc Li et al. (2025a) probes action confusions between similar clips. We report accuracy on all claims, on positive claims, and on negative claims, together with its official action hallucination score, which counts a question only if every clip is answered correctly. General capability is measured by the official multiple-choice accuracy of MMVU Zhao et al. (2025), Video-MMMU Hu et al. (2026), and MMBench Liu et al. (2024). All models are evaluated by us with 32 frames and greedy decoding.

A.3 Additional Results

VideoHallucer by category. Table 1 reports three of the five categories of VideoHallucer, and Table 5 gives all five. CF-GRPO is the best trained model on object relation and external nonfactual questions, at 64.50 and 58.75 percent. The negative questions of these two categories describe a relation or an event that is plausible for the scene but absent from the video, which is how the absent event e_{abs} of Section 3.2 is chosen. On the other three categories CF-GRPO stays within two points of the best trained model. Every trained model stays below 30 percent on external factual questions, because their negative questions state facts that hold in the world but not in the video, and the language prior confirms such facts.

Number of evaluation frames. All results in Section 4 use 32 evaluation frames. Table 6 repeats the evaluation with 8 and 16 frames to test whether the comparison depends on this choice. At every frame count CF-GRPO leads every other trained model on EventHallusion and VideoHallucer and matches GRPO on VidHalluc, as it does at 32 frames. Its accuracy varies by about one point at most across the three frame counts, so the comparison does not depend on number of evaluation frames.

A.4 Data Construction

Base samples. The base set is built from the dense captions of LLaVA-Video-178K. Given a caption, Qwen2.5-VL-7B-Instruct proposes up to three minimal edits, each changing one visually verifiable attribute of one sentence so that the sentence becomes false for the video (Appendix A.6). An edit is kept only if the edited span is a verbatim substring of the caption of at most twelve words, is not a reordering of the same words, and contains no hedging or subjective wording, which leaves $3{,}176$ pairs of an original sentence, labeled yes, and its edited version, labeled no. The cold-start model then answers each member six times on the video with the training prompt at temperature 1.0, and a pair is kept only if at least one member is answered correctly in some but not all samples, which removes edited sentences that remain true of the video and pairs that the model already answers with certainty; $2{,}220$ pairs remain. Requiring evidence keys for both members leaves $1{,}529$ pairs, that is, $3{,}058$ samples. The evidence keys of a negative sample are the content words of the original attribute, which the edited statement does not contain, and those of a positive sample are the content words of another true attribute of the same video that its statement does not contain.

Counterfactual pairs. For each of $1{,}800$ captions, the base model proposes an event triple, the dominant event e_{main}, a secondary event e_{sub}, and an absent event e_{abs} that is plausible for the scene. A second pass of the base model with the video verifies that e_{main} is dominant, that e_{sub} occurs, and that e_{abs} does not occur while the entities it names are present. $1{,}211$ triples pass, and 500 of their videos are used. Each video yields the two pairs of Equation 9 under the templates *Are people [event] in the video?* for $O(e)$ and *Is [event] the main activity in the video?* for $D(e)$, which gives $2{,}000$ samples. The evidence keys of a pair sample are the words of the queried event. The pool of $5{,}058$ samples holds $2{,}529$ with label yes and $2{,}529$ with label no.

Output format. Every output follows a template of four stages fixed at the cold start, observation, evidence, reasoning, and answer, and the format gate f_i of Equation 15 is 1 only if all four are present and the output reaches a minimum length. The evidence check reads stages before the answer.

A.5 Design of the Counterfactual Pairs

Proposition 2 requires two prompts with opposite labels that the prior answers alike. Replacing the queried event while keeping the video and the template gives opposite labels for any video, and choosing both events to be plausible for the scene leaves the prior no basis to prefer one member.

The alternatives fail one of the two requirements. Editing the video, as Video-R1 and CRPO do, leaves the label unchanged unless the question concerns temporal order. The prior alone rejects an unrelated event, so it answers the two members differently and the negative member can be answered without the video. A negated statement changes the text rather than the event, so the two members can be distinguished by their text. The absent event is therefore proposed from the caption alone and kept only if no sentence of the caption states or entails it while the entities it names are present, so the labels follow the annotated captions rather than the model's own reading of the video.

The two templates address two ways in which the prior confirms a statement that the video does not support. The template of $O(e)$ pairs an event that occurs with one that does not, the failure probed by the object relation and external nonfactual categories of Table 5. The template of $D(e)$ pairs the dominant event with a secondary event that also occurs, so the model must judge which event the video is about rather than whether an event is present. Since e_{sub} is positive under O and negative under D, the event cannot serve as a label, and its evidence keys appear in one positive and one negative sample, so stating them is rewarded only together with the correct answer. Using e_{sub} rather than a paraphrase of the dominant event as the positive member of the occurrence pair also avoids rewarding restatements, which the variant of Section 4.3 shows to lower accuracy.

A.6 Prompts for Data Construction

All prompts below are answered by Qwen2.5-VL-7B-Instruct from the caption alone, with the output constrained to JSON. Generation uses temperature 0.3, and the verification of the pairs uses greedy decoding. The base-set edits are filtered by rules and by a pass-rate prescreen that uses the training prompt below, not by a separate verification prompt. The prompts are reproduced as used; only sentence punctuation of the training prompt is added for readability, and *salience twin* is the internal name of a counterfactual pair. The fields `main_action`, `subsidiary`, and `plausible_nonevent` are e_{main}, e_{sub}, and e_{abs} of Section 3.2; the field `paraphrase` is used only by the ablation variant of Section 4.3.

CF-GRPO: Training Prompt

User. {video}
Based on the video, is the following description accurate? "{statement}"
Please carefully analyze the pictures (or videos) and problems according to the following requirements. In <prethink> </prethink> tags, carefully analyze the problem and briefly explain the steps to explain the problem and the key thinking direction of reasoning the problem. In <caption> </caption> tags, Please describe the image carefully, paying special attention to the details related to the problem and the reasoning direction of solving the problem. In <think> </think> tags, outline a step-by-step thought process you would use to solve the problem based on the image. In <answer> </answer> tags, give the final answer in a direct format, and it must match the correct answer exactly. Please sort out the output in the format of '<prethink>...</prethink> <caption>...</caption> <think>...</think> <answer>...</answer>' according to the above requirements. Please provide only Yes or No within the <answer> </answer> tags.

Statement. For a base sample, a sentence of the caption or its minimal edit. For a pair sample, `The video shows {e}.` under $O(e)$ or `The main activity in the video is {e}.` under $D(e)$.

CF-GRPO: Counterfactual Pair Generation Prompt

System. You analyze a TRUE dense caption of a video and extract a salience structure for building yes/no training questions. Return JSON with exactly these fields:

- `main_action`: the single DOMINANT action or event of the video, a short verb phrase grounded in the caption (the thing the video is mainly about).
- `paraphrase`: the SAME main action reworded with different verbs or synonyms, still an accurate description (e.g. 'slicing an onion' → 'cutting an onion into pieces').
- `subsidiary`: an action that DOES occur in the caption but is clearly secondary or preparatory, NOT the main action (e.g. 'washing the knife' in a cooking video).
- `plausible_nonevent`: an action that does NOT occur anywhere in the caption but is plausible given the entities and scene (e.g. 'the man plays with the dog' when a man and a dog are present but never interact). It must be checkable and clearly absent.

All four must be short verb phrases under 12 words. Do not copy whole sentences.

User. Example caption: """The video begins with a man in a red jacket kneeling on the floor next to a wooden chair, with a dog resting nearby. He picks up a screwdriver and tightens the screws on the chair. After tightening the screws, he flips the chair upright and sits on it to test the stability."""
Good output: {"main_action": "repairing a wooden chair with a screwdriver", "paraphrase": "fixing a chair by tightening its screws", "subsidiary": "sitting on the chair", "plausible_nonevent": "playing with the dog"}
Now extract the salience structure for this caption.
CORRECT caption: """{caption}"""

CF-GRPO: Counterfactual Pair Verification Prompt

System. You are a strict checker for salience-twin extraction. The caption is the TRUE description of a video. Check four conditions:

- c1 main_ok: main_action is truly the dominant action of the caption.
- c2 para_ok: paraphrase describes the SAME action as main_action (synonym-level, no new claims) and is accurate for the caption.
- c3 sub_ok: subsidiary DOES occur in the caption AND is clearly not the dominant action (if it could be argued to be the main action, answer false).
- c4 non_ok: plausible_nonevent does NOT occur anywhere in the caption (no sentence states or entails it), yet its entities are present in the scene.

Write a one-sentence reason first, then the four booleans.

User. caption: """{caption}"""
main_action: $\{e_{\text{main}}\}$ paraphrase: {paraphrase} subsidiary: $\{e_{\text{sub}}\}$ plausible_nonevent: $\{e_{\text{abs}}\}$
Check c1-c4.

A triple is kept only if all four conditions hold.

CF-GRPO: Base Sample Edit Prompt

System. You build HARD NEGATIVE captions to train a video-language model to make fine-grained distinctions. You are given a CORRECT caption of a video. Produce minimal edits: each changes EXACTLY ONE short, visually-verifiable attribute so the caption becomes FALSE for the video, while staying fluent and plausible.
Rules:

- original_span MUST be copied VERBATIM from the caption (an exact substring).
- Change as few words as possible; never rewrite the whole sentence.
- The attribute MUST require watching the video (count, action/verb, order of events, motion direction, who/what performs the action, magnitude, spatial relation, or a concrete object attribute). NOT world knowledge.
- The edit must be a REAL visual difference, not a synonym/paraphrase.
- PREFER temporal/dynamic attributes (action_type, action_frequency, event_order, motion_direction, motion_magnitude, action_effector) over static ones.
- Keep each original_span SHORT (¡= ˜12 words).

User. Example caption: """The video begins with a man in a red jacket kneeling on the floor. He picks up a screwdriver with his right hand and tightens the screws on a wooden chair. After tightening the screws, he flips the chair upright and sits on it to test the stability."""
Good output: {"edits": [{"category": "action_effector", "original_span": "picks up a screwdriver with his right hand", "edited_span": "picks up a screwdriver with his left hand", "focus_question": "Which hand does the man use to pick up the screwdriver?", "chosen_answer": "his right hand", "rejected_answer": "his left hand"}, {"category": "event_order", "original_span": "After tightening the screws, he flips the chair upright", "edited_span": "Before tightening the screws, he flips the chair upright", "focus_question": "Does the man flip the chair before or after tightening the screws?", "chosen_answer": "after tightening the screws", "rejected_answer": "before tightening the screws"}, {"category": "action_type", "original_span": "sits on it to test the stability", "edited_span": "stands on it to test the stability", "focus_question": "What does

```
the man do on the chair to test its stability?", "chosen_answer":
"sits on it", "rejected_answer":  "stands on it"}]}
```

Bad edits (do NOT produce): synonym swaps (”picks up” → ”grabs”), world knowledge (”screwdriver” → ”hammer” when no hammer visible is fine ONLY if the tool is clearly visible), whole-sentence rewrites.
Now produce up to 3 minimal edits for this caption.
CORRECT caption: """{caption}""" An edit is kept if its span is a verbatim substring of the caption of at most twelve words, is not a reordering of the same words, and contains no hedging or subjective wording. The kept pairs are then screened by the pass-rate prescreen of Appendix A.4, which uses the training prompt above.

A.7 More Qualitative Examples

Figures 4 to 8 show five negative questions of VideoHallucer that GRPO and OMNEX-VL both confirm and CF-GRPO rejects. Their ten wrong answers fall into three groups: answering from what is typical for the scene (Figures 4 and 8), stating the relation that the video shows and then contradicting it in the answer, the case that the penalty w_c of Equation 15 targets (Figures 5 and 6), and misreading the video. In every case the reasoning of CF-GRPO states what the video shows and its answer follows it.



Figure 4: **Semantic detail.** The man holds a microphone, not a pen. GRPO takes the object near the microphone for a pen, and OMNEX-VL answers from the common practice of holding a pen during a presentation. CF-GRPO names the microphone and finds no pen.

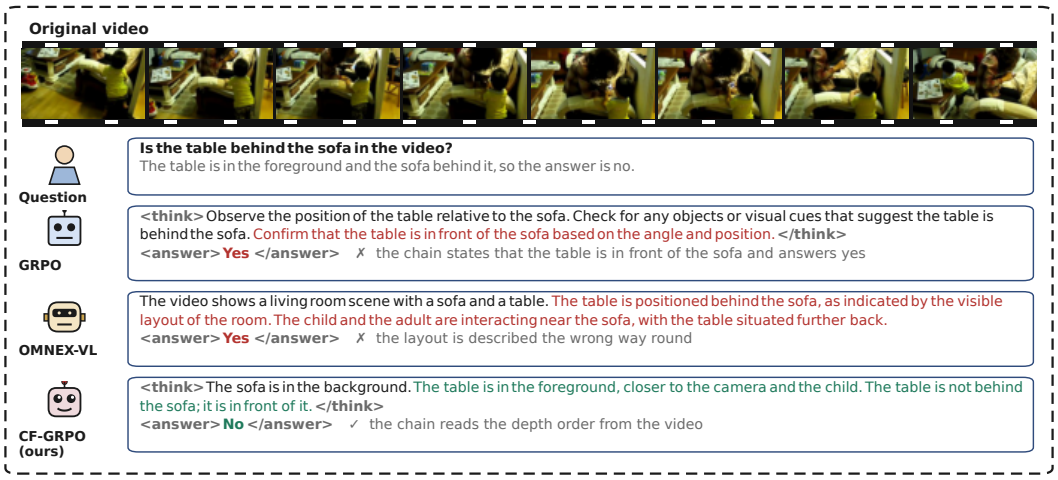


Figure 5: **Object relation.** The table stands in the foreground and the sofa behind it. GRPO states in its reasoning that the table is in front of the sofa and still confirms that it is behind, the case that the evidence-consistency reward of Section 3.4 penalizes. OMNEX-VL describes the layout the wrong way round, and CF-GRPO reads the depth order from the video.

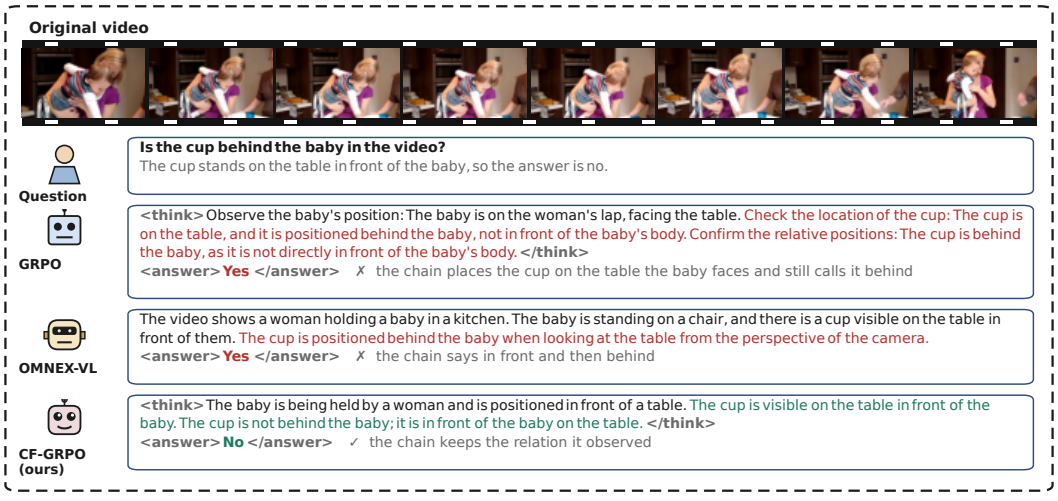


Figure 6: **Object relation.** The cup stands on the table in front of the baby. GRPO and OMNEX-VL both place the cup on the table that the baby faces and still confirm that it is behind the baby, whereas CF-GRPO keeps the relation it observes.

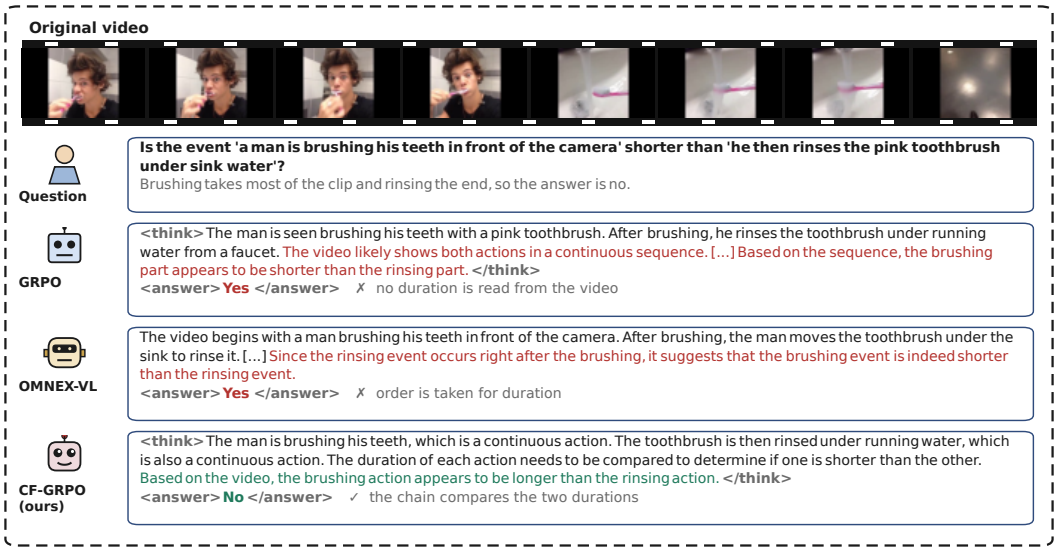


Figure 7: **Temporal.** Brushing takes most of the clip and rinsing only the end. GRPO and OMNEX-VL both infer the durations from the order of the two actions, whereas CF-GRPO compares the durations in the video.

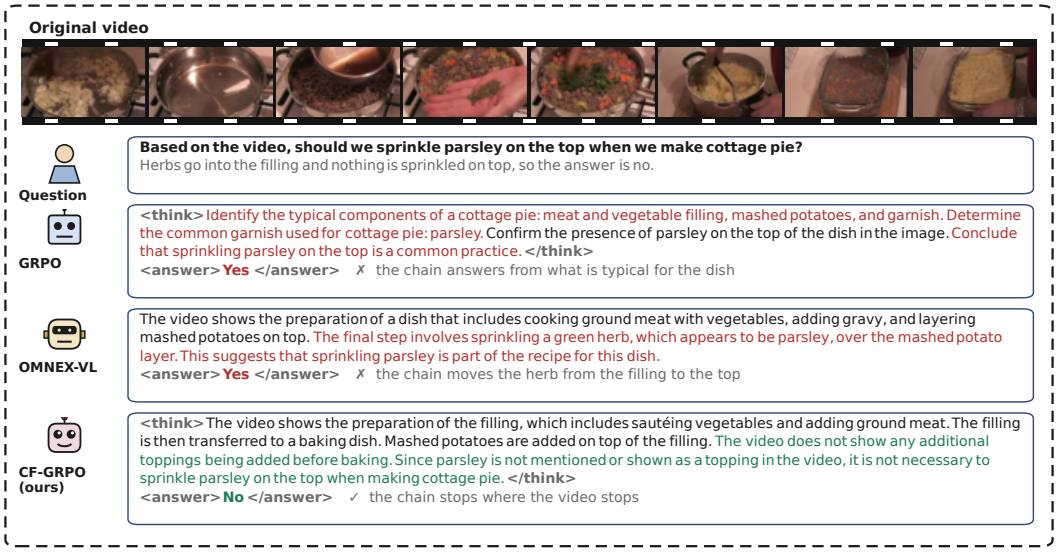


Figure 8: **External factual.** Herbs go into the filling and nothing is sprinkled on top. GRPO answers from the usual garnish of the dish, and OMNEX-VL moves the herb from the filling to the top. CF-GRPO notes that the video shows no topping added before baking.