Perception Without Engagement: Dissecting the Causal Discovery Deficit in LMMs

Jiafeng Liang[1]*, Zhihao Zhu[1]*, Zihan Zhang[5]*, Baoqi Ren[1], Shixin Jiang[1], Runxuan Liu[1], Tao Ren[4], Ming Liu[1,2] , See-Kiong Ng[3], Bing Qin[1,2]†
[1]Harbin Institute of Technology, [2]Pengcheng Laboratory, [3]National University of Singapore,
[4]Peking University [5]Harvard University
{jfliang, zhzhu, mliu}@ir.hit.edu.cn

Abstract

Although Large Multimodal Models (LMMs) have achieved strong performance on general video understanding, their susceptibility to textual prior shortcuts during causal discovery has been recognized as a critical deficit. The underlying mechanisms of this phenomenon remain incompletely understood, as existing benchmarks only measure response accuracy without revealing the sources and extent of the deficit. We introduce **ProCauEval**, a perturbation-based evaluation protocol that shifts from outcome assessment to mechanism diagnosis, probing causal discovery through five controlled configurations that systematically manipulate visual and textual modalities to decompose their respective contributions to model behavior and dissect the failure modes. Evaluating 17 mainstream LMMs, we find that models faithfully perceive video content yet systematically underexploit it during causal reasoning. We further observe that stronger post-training amplifies rather than mitigates textual prior reliance, and that higher performance correlates with greater fragility under perturbation. To address these, we propose **A**nti-**D**istillation **P**olicy **O**ptimization (**ADPO**), a reinforcement learning framework built on negative teacher alignment, which augments GRPO by explicitly pushing the policy away from a prior-only counterfactual teacher induced by visual corruption. Specifically, ADPO maximizes the divergence between the policy distributions conditioned on the original and visually corrupted inputs, thereby forcing the model to ground its reasoning in visual evidence rather than textual shortcuts. Extensive experiments show that ADPO improves visual engagement without sacrificing fundamental comprehension, thus offering a preliminary step toward reliable causal discovery.

1 Introduction

Large Multimodal Models (LMMs) [1–3] have achieved remarkable progress across a wide range of video understanding tasks such as action recognition [4], visual question answering [5], and scene description [6]. Yet in real-world scenarios, video understanding demands more than perceiving what is visible, requiring instead the ability to reason about why observed events occur, such as inferring why a vehicle brakes suddenly in autonomous driving or what triggers an abnormal tissue response in medical video analysis. These settings share a common demand for causal discovery [7], namely identifying directed cause-effect relationships between events, a qualitatively harder challenge than the perceptual recognition tasks dominating current benchmarks. Given this gap, whether LMMs genuinely possess robust causal discovery capability remains an open and pressing question.

Recently, causal video benchmarks [8, 7] have made notable progress, most evaluate causal discovery exclusively through outcome accuracy, measuring *how well* a model performs without investigating *why* it fails or *to what extent* each failure source contributes to the overall deficit. Some benchmarks [9,

* Equal contribution
† Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

10] have recognized the risk of statistical shortcut leveraging, where models arrive at correct answers by relying on superficial cues such as linguistic patterns or knowledge biases rather than genuinely reasoning over visual content. To mitigate this, they attempt to remove exploitable cues from the evaluation data, whether by filtering text-solvable questions or avoiding predictable answer formats. However, such a defensive strategy faces an inevitable arms race, as increasingly capable models inevitably discover subtler statistical regularities that benchmark designers did not anticipate.

To address this gap, we introduce **ProCauEval**, a perturbation-based evaluation protocol that shifts the paradigm from outcome evaluation to mechanism dissection. Rather than eliminating shortcuts, ProCauEval deliberately injects them through controlled perturbations and examines whether models still ground their reasoning in visual evidence. Specifically, ProCauEval defines five controlled configurations, namely Veridicality, Plausibility, Incongruence, Deprivation, and Diagnosis, which jointly perturb both visual and textual modalities, enabling a causal decomposition of each modality's contribution to model behavior. Complemented by five evaluation metrics that capture flip, non-flip, and perceptual fidelity beyond answer accuracy, the protocol provides a fine-grained characterization of the extent and mechanisms of failures in model causal discovery.

Using ProCauEval, we conduct a systematic evaluation of 17 mainstream LMMs spanning instruction-tuned, reasoning-oriented, and large-parameter variants, and arrive at a set of findings. First, beyond the known risk of textual shortcut exploitation, our controlled decomposition pinpoints the failure occurring at the reasoning stage rather than perception, as models faithfully comprehend video content yet systematically discard it when inferring causal relations. Second, reasoning-oriented post-training [11] amplifies rather than mitigates this dependence on textual priors and may even render visual engagement counterproductive, as models perceive video content yet are misled rather than informed by it during causal reasoning. Third, higher baseline accuracy correlates strongly with greater fragility under perturbation, revealing that top-performing models are precisely those whose apparent causal discovery is most unsubstantiated. These findings collectively locate the core deficit not in visual perception but in cross-modal integration and reasoning, where visual signals are systematically suppressed by textual priors within the thinking process.

Building on this dissection, we further propose **A**nti-**D**istillation **P**olicy **O**ptimization (**ADPO**), a reinforcement learning framework that directly targets the prior-dependent failure mode exposed by ProCauEval. In contrast to conventional distillation [12], which pulls a student toward a teacher, ADPO pushes the policy away from a degenerate prior-only teacher that has been deprived of visual input. Specifically, built upon Group Relative Policy Optimization (GRPO), ADPO introduces a perception-sensitivity regularizer that maximizes the divergence between the policy distributions conditioned on the original video and on a visually corrupted counterpart, thereby forcing the model to ground its reasoning in visual evidence. As shown in Figure 1, ADPO significantly increases the model's likelihood preference for correct answers over prior shortcut, steering models toward robust causal discovery more firmly grounded in visual evidence. Our main contributions are as follows:

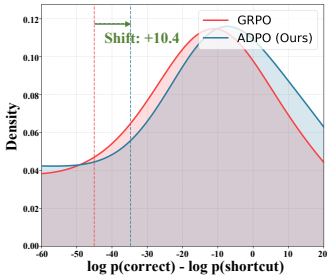


Figure 1: Correct and shortcut output likelihood distribution gap between GRPO and our proposed ADPO.

- We design ProCauEval, a perturbation-based evaluation protocol with five controlled configurations, shifting causal evaluation from outcome accuracy to mechanism diagnosis.
- We evaluate 17 mainstream LMMs and reveal that their causal discovery is largely spurious, with the failure located in cross-modal integration and reasoning rather than perception.
- We propose Anti-Distillation Policy Optimization (ADPO), a reinforcement learning framework that augments GRPO with a perception-sensitivity regularizer to improve model visual engagement.

The source code, model weights, and data will be made openly available to facilitate future research.

2 Related Work

Causal Discovery Benchmarks. Evaluating causal discovery in video models has attracted growing attention [5, 13, 14]. MECD [7] formulates event-level causal graph discovery over videos, and CausalStep [8] extends this to stepwise causal chain reasoning. CausalVQA [9] and VCRBench [10] further probe causal capabilities through physically grounded and long-form video settings, respectively. A recurring concern in these benchmarks is shortcut exploitation, where models rely on

statistical cues rather than visual evidence. Existing fixes such as filtering text-solvable questions, designing harder distractors, or balancing answers are defensive patches that struggle to prevent stronger models from capturing subtle regularities. Our PROCAUEVAL instead injects controlled perturbations to diagnose failure mechanisms, shifting from outcome evaluation to failure dissection.

Group Relative Policy Optimization for Video Reasoning. Group Relative Policy Optimization (GRPO) [11] has become the dominant reinforcement learning framework for post-training LMMs [15–19]. For video reasoning, most existing methods simply extend GRPO with visual inputs and directly transfer the algorithm to video tasks [20–23] or optionally augmenting it with auxiliary task-specific scalar rewards [24–26]. Although these methods improve chain-of-thought quality, our analysis indicates that they paradoxically amplify reliance on textual priors, as the verifier rewards answer correctness without distinguishing whether it arises from visual grounding or textual co-occurrence. Departing from conventional distillation that pulls a student toward a positive teacher [12], our ADPO inverts this paradigm by treating a visually-deprived policy as a negative teacher and repelling the grounded distribution away from it via asymmetric stop-gradient KL.

3 Causal Evaluation Protocol (PROCAUEVAL)

We propose a causal evaluation protocol (PROCAUEVAL), covering framework, data, and metrics, shifting the evaluation paradigm from outcome accuracy to mechanism dissection. The core idea is to actively inject controlled perturbations and examine whether model judgments shift in response, thereby quantitatively characterizing the sources and extent of failures in causal discovery.

3.1 Causal Discovery in Video Comprehension

Causal discovery [7] refers to the task of identifying causal relationships among chronologically ordered events within a video, with the goal of constructing a structured, event-level causal diagram that explains why the final event occurred. Formally, given a video V accompanied by an event set $E = \left\{e_1^{ca}, e_2^{ca}, ..., e_{n-1}^{ca}, e_n^{ef}\right\}$ where all events correspond to occurrences within the video. The final event $E_e = \left\{e_n^{ef}\right\}$ is designated as the effect event, while the remaining candidate events are partitioned into two subsets based on a binary label indicating whether an event constitutes a cause. Specifically, the cause event set $E_c = \left\{e_j^{ca} \mid \text{label}(e_j^{ca}) = 1\right\}$ contains events labeled as causes, whereas the non-cause event set $E_o = \{e_k^{ca} \mid \text{label}(e_k^{ca}) = 0\}$ contains events negatively labeled.

3.2 Evaluation Framework

To conduct a comprehensive ablation and comparative analysis, we systematically design five distinct data pairing configurations. Specifically, by selectively modifying the input content of a given modality, we are able to progressively disentangle the causal influence of each modality on the LMMs, and identify the underlying reasons for the failure. See Appendix B for more details.

3.2.1 Data Elements

We first construct the data elements (shown in Figure 2 (b)), which serve as the fundamental building blocks for the evaluation configurations described below. The data elements are defined as:

- ***a. Real Events*:** Events that actually occurred in the video and are described in textual form.
- ***b. Real Video*:** Video that authentically contain multiple causal events (i.e., video clip form).
- ***c. Fake Events*:** The textual event set is constructed by inserting fabricated cause events $E_f = \{e_l^{ca}\}$ into the original event sequence. We use GPT-5 [27] to generate candidate fabricated causes and retain those with higher language-only log-probability than the ground-truth causes, making them statistically plausible distractors rather than model-specific adversarial examples.
- ***d. Fake Video*:** Constructed by replacing cause event clips in real videos with fabricated ones generated by Google VEO-3 [28], conditioned on the corresponding cause event text descriptions. By design, generation is conditioned solely on event semantics, yielding clips that are semantically faithful yet scene-inconsistent with surrounding context, a deliberate property probing cross-modal integration. Low-quality clips are filtered via CLIP similarity [29] thresholding.
- ***e. Blank Video*:** All event clips are replaced with blank segments of white frames.

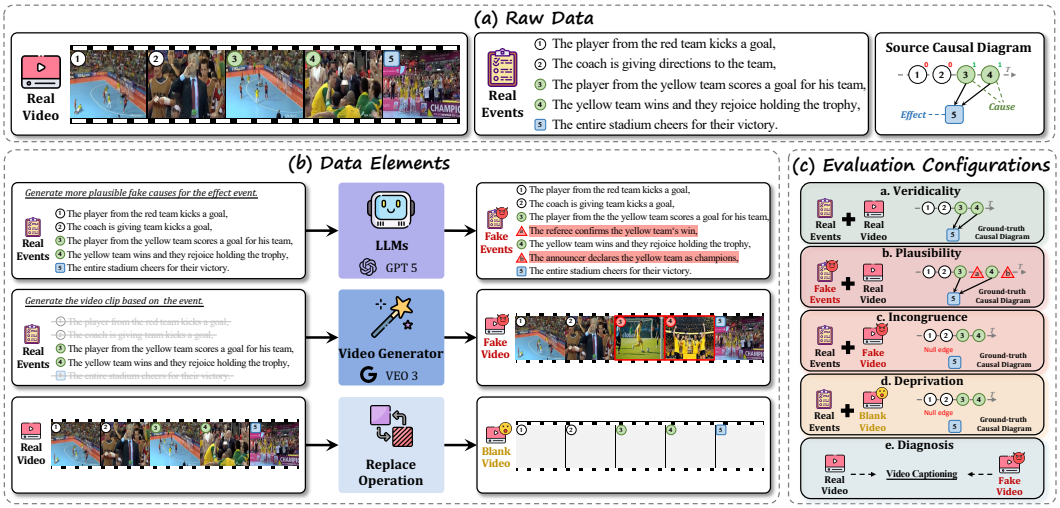


Figure 2: Overview of PROCAUEVAL framework covering data elements and configurations.

3.2.2 Evaluation Configurations

To comprehensively assess the model's causal discovery capability, underlying mechanisms and potential failure modes, we establish five evaluation configurations (shown in Figure 2 (c)).

- ***a. `Veridicality` (Real Events + Real Video)*:** Primarily serves to verify whether the model possesses causal capability. Rather than examining the source or reliability of the model's causal discovery, it focuses solely on assessing whether correct judgments can be made.
- ***b. `Plausibility` (Fake Events + Real Video)*:** This configuration investigates whether the model relies on statistical correlations derived from textual priors alone (i.e., pre-trained knowledge rather than genuinely reasoning over the provided content) when performing causal discovery.
- ***c. `Incongruence` (Real Events + Fake Video)*:** To examine whether the model performs causal discovery by jointly reasoning over the full visual context and aligning it with the textual events.
- ***d. `Deprivation` (Real Events + Blank Video)*:** This configuration introduces a single-variable difference in the video modality relative to `Veridicality`, enabling controlled comparison.
- ***e. `Diagnosis` (Real and Fake Video Captioning)*:** To assess general video understanding independently of causal discovery, captions are generated from both original and fabricated cause videos clips. This rules out the possibility that observed failures in causal discovery reflect an inability to comprehend video content rather than genuine deficiencies in causal reasoning.

3.3 Data Statistics

Our evaluation data is built upon the causal discovery benchmark MECD [7]. We collect 500 videos and automatically construct 2,000 QA pairs based on the original annotations. Each video is associated with four distinct evaluation settings based on inconsistency perturbations. Furthermore, to ensure the comprehensiveness of the evaluation, we deliberately select videos of varying durations. Details are presented in Table 1.

Table 1: Data statistics

Category	Size
Video Sources	500
- Avg events per Video	5.1
- Max Duration	536.9s
- Min Duration	8.9s
- Avg Duration	132.8s
Questions per Video	4
Total Samples	2000

3.4 Evaluation Metrics

In this section, we introduce the five proposed metrics used in our evaluation. We begin with the general metrics that are applicable to most evaluation configurations:

Average Score (R-Avg): R-Avg measures the average alignment between predicted cause events and cause events in the source causal diagram. For sample i, given the set of cause events E_c from the

source causal diagram and the set of predicted cause events $\hat{E}_c$, R-Avg is defined as:

$$\text{R-Avg} = \frac{1}{D}\sum_{i=1}^{D}\frac{\left|E_c \cap \hat{E}_c\right|}{|E_c|}, \tag{1}$$

where D denotes the total number of test samples, $|\cdot|$ denotes the cardinality of a set. Furthermore, we introduce four additional proposed metrics, each tailored to specific evaluation configurations:

Accuracy (Acc): Acc is specifically designed for `Incongruence` and `Deprivation`, in which the ground-truth causal diagram contains no causal edges. To accommodate this, we introduce an additional candidate option $E_n = \{\text{“Based on this video, none of them”}\}$, where selecting this option is the only correct answer, formulated as:

$$\text{Acc} = \frac{1}{D}\sum_{i=1}^{D}\mathbf{1}\left[\hat{E}_c = E_n\right]. \tag{2}$$

Flip Rate (FR): FR is proposed for configuration `Plausibility` to assess how susceptible the model is to statistically-biased textual perturbations. Specifically, FR measures the proportion of responses that were correct under configuration `Veridicality` but were flipped to match the fabricated events introduced in configuration `Plausibility`, formulated as:

$$\text{FR} = \frac{\sum_{i=1}^{D}\mathbf{1}_{\mathbf{V}}\left[\frac{\left|E_c\cap\hat{E}_c\right|}{|E_c|}=1\right]\cdot\mathbf{1}_{\mathbf{P}}\left[\frac{\left|E_f\cap\hat{E}_c\right|}{|E_c|}=1\right]}{\sum_{i=1}^{D}\mathbf{1}_{\mathbf{V}}\left[\frac{\left|E_c\cap\hat{E}_c\right|}{|E_c|}=1\right]}, \tag{3}$$

where $\mathbf{1}_{\mathbf{V}}[\cdot]$ and $\mathbf{1}_{\mathbf{P}}[\cdot]$ are indicator functions under configurations `Veridicality` and `Plausibility`, respectively. A higher FR indicates that the model is more easily misled into selecting fabricated events over the original correct causes.

Non-Flip Rate (NFR): When video content is replaced, an ideal model should update its causal judgments accordingly. Specifically, NFR measures the proportion of responses that were correct under `Veridicality` and remained unchanged after the video was manipulated in configurations `Incongruence` or `Deprivation`.

$$\text{NFR} = \frac{\sum_{i=1}^{D}\mathbf{1}_{\mathbf{V}}\left[\frac{\left|E_c\cap\hat{E}_c\right|}{|E_c|}=1\right]\cdot\mathbf{1}_{\mathbf{I,D}}\left[\frac{\left|E_c\cap\hat{E}_c\right|}{|E_c|}=1\right]}{\sum_{i=1}^{D}\mathbf{1}_{\mathbf{V}}\left[\frac{\left|E_c\cap\hat{E}_c\right|}{|E_c|}=1\right]}, \tag{4}$$

where $\mathbf{1}_{\mathbf{I,D}}[\cdot]$ means indicator functions under configurations `Incongruence` or `Deprivation`. A higher NFR exposes the model's failure to leverage visual information for causal reasoning.

GPT Score: GPT Score is introduced for configuration `Diagnosis` to evaluate the fine-grained quality of video comprehension. We design a detailed prompt that assesses each caption from multiple dimensions, assigning score on scale of 1 to 10. See Appendix B for evaluation prompt.

4 Evaluations

4.1 Models

We conduct a comprehensive evaluation of most existing 17 mainstream LMMs, encompassing both standard instruction-tuned models and reasoning-oriented thinking models. Additionally, we extend our evaluation to include models of larger parameter scales. See Appendix B for more details.

4.2 Results and Analysis

Videos are Substantially Underused in Causal Discovery. A central finding from Table 2 is that video information is largely bypassed during causal discovery. The most direct evidence comes from the `Plausibility`, in which models that genuinely ground their judgments in visual content would

Table 2: Evaluation results of three types of LMMs under four configurations.

Model	Size	Veridicality	Plausibility		Incongruence			Deprivation		
		R-Avg ↑	R-Avg ↑	FR ↓	R-Avg ↓	Acc ↑	NFR ↓	R-Avg ↓	Acc ↑	NFR ↓
Standard Models										
LLaVA-OneVision [1]	7B	20.4	2.6	24.7	20.6	28.2	53.3	23.8	34.4	33.8
Qwen3-VL [2]	8B	36.4	3.9	51.7	31.2	13.2	54.1	36.7	24.6	52.5
InternVL-3.5 [3]	8B	34.1	2.7	30.9	32.6	17.0	43.1	50.3	28.8	51.2
MiMo-VL-SFT [26]	7B	34.5	7.9	57.9	33.2	6.8	63.2	35.8	17.6	47.4
GPT-5.1 [27]	-	25.8	1.0	20.8	20.5	28.6	40.6	29.3	23.4	43.4
Thinking Models										
Video-R1 [24]	7B	27.9	4.9	22.3	31.6	16.4	52.4	32.8	26.2	45.6
VideoChat-R1.5 [25]	7B	37.1	11.9	29.1	36.6	17.6	60.9	30.3	19.2	50.4
Keye-1.5 [23]	8B	47.8	3.6	70.0	51.2	14.0	78.0	49.2	12.0	73.3
LongVILA-R1 [20]	7B	20.6	5.2	23.0	18.9	7.4	33.8	25.3	11.8	46.0
GLM-4.1V-Thinking [30]	8B	37.7	1.1	36.7	37.3	9.8	65.5	43.9	18.2	69.1
Qwen3-VL-Thinking [2]	8B	47.6	3.8	72.9	50.1	8.0	79.3	50.2	8.8	79.3
OneThinker [22]	8B	31.2	2.5	36.8	33.1	4.2	50.9	28.3	20.4	46.9
MiMo-VL-RL [26]	7B	30.0	6.2	37.8	34.1	10.2	52.4	38.3	17.0	56.1
GPT-5.1-Thinking [27]	-	33.9	0.4	31.9	38.3	11.0	70.2	39.9	11.4	75.2
Larger Models										
Qwen3-VL [2]	32B	37.4	2.6	37.1	35.0	13.4	68.6	39.3	17.2	62.1
InternVL-3.5 [3]	38B	26.8	1.6	30.2	29.4	17.4	58.2	43.1	10.4	59.3
Qwen3-VL-Thinking [2]	32B	48.0	3.2	55.8	42.4	8.2	63.8	49.3	22.0	65.0

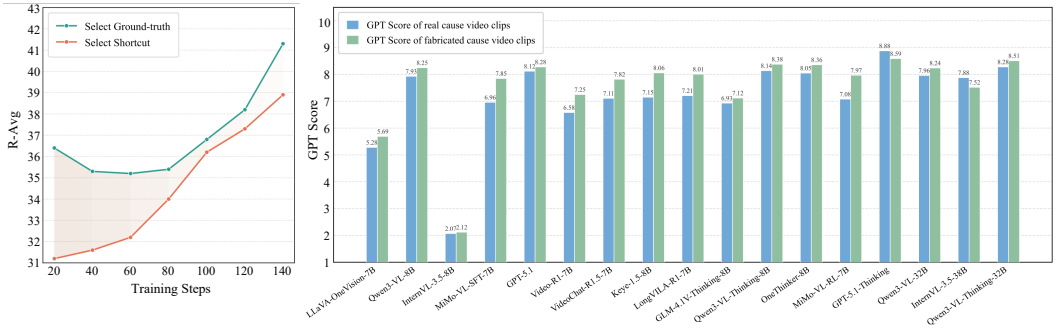


Figure 3: (a) Selection trends of the model toward correct and shortcut answers during GRPO training. (b) GPT Scores of LMMs' perception capability on real and fabricated cause video clips.

be expected to maintain a stable R-Avg. However, the mean R-Avg across all 17 LMMs collapses from 33.9% under `Veridicality` to merely 2.4%, corresponding to an average relative drop of 88.7%. This conclusion is reinforced from the opposite direction by the `Incongruence` and `Deprivation`, where replacing real videos with fabricated content yields a mean R-Avg difference of only 3.7%, and 15 out of 17 models attain an even higher R-Avg when videos are substituted with blank frames. Moreover, this deficit is not reliably alleviated by scaling up parameters, as the three scaling pairs in the table exhibit divergent and even inverse trends. Taken together, these findings suggest that videos do not serve as an effective source of evidence for causal discovery, and models tend to fall back on textual priors, producing judgments that are largely ungrounded.

Fundamental Comprehension and Causal Discovery are Decoupled. A natural concern is whether these failures simply reflect an inability to comprehend video content, a hypothesis directly refuted by the `Diagnosis`. As shown in Figure 3 (b), most LMMs produce captions that faithfully and comprehensively describe both real and fabricated videos, achieving average GPT Scores of 7.3, indicating reliable perception at the visual level. The contrast with their behavior under `Incongruence` is striking, as exemplified by Qwen3-VL-8B-Thinking [2], which attains a GPT Score of 8.14 on real videos yet only 8.0% Acc under and a 79.3% NFR under `Incongruence`. This decoupling between perception and evidence utilization supports our central claim that the causal discovery deficit in LMMs is not a perceptual failure but one of cross-modal integration, in which visual signals are systematically suppressed by textual priors.

Post-training Induces an Inverse Scaling between Reasoning Capacity and Visual Engagement. Reasoning-oriented RL with explicit chain-of-thought (CoT), although the dominant post-training paradigm for reasoning, fails to mitigate and even exacerbates the reliance on textual priors. As shown in Table 2, thinking variants attain marginally higher mean R-Avg under `Veridicality`

than standard variants, yet exhibit substantially higher NFR and lower Acc. A controlled evaluation on Qwen3-VL-8B with GRPO further confirms this trend (shown in Figure 3 (a)), as the R-Avg gap between `Veridicality` and `Incongruence` narrows progressively across training steps. Moreover, several thinking models achieve higher R-Avg under `Incongruence` than under `Veridicality`, a pattern largely absent in standard models. One plausible explanation is that the extended reasoning chains of thinking models perceive and incorporate rich visual details from real videos, yet lack the capacity to correctly leverage them for causal judgment. When replaced with scene-inconsistent fabricated content, the visual stream signals a cross-modal conflict, prompting models to fall back to textual priors, paradoxically achieving higher accuracy. This improvement is essentially ungrounded, as the model is not reasoning better but has disengaged from the video.

Higher Accuracy Comes with Greater Fragility. One would intuitively expect that LMMs with higher causal discovery accuracy possess stronger causal capabilities and thus exhibit greater robustness, yet our results reveal precisely the opposite pattern. As shown in Table 3, the Pearson correlation between R-Avg under `Veridicality` and FR under `Plausibility` reaches $r = 0.832$ across all 17 LMMs, and similarly strong correlations are observed for NFR under `Incongruence` and `Deprivation`, indicating that higher baseline accuracy is consistently associated with greater susceptibility to textual perturbation. This paradox is most clearly illustrated by the top-performing LMMs under `Veridicality`, whose correct predictions are largely flipped to fabricated events under textual perturbation and remain nearly unchanged even when the video is entirely replaced. Their high accuracy is therefore not a product of grounded multimodal causal discovery, but rather reflects well-calibrated textual priors that happen to align with the pre-training pattern.

Table 3: Cross-metric correlation analysis revealing the accuracy-fragility paradox. `V`, `P`, `I`, and `D` denote the `Veridicality`, `Plausibility`, `Incongruence`, and `Deprivation` configurations, respectively.

Metric Pair	**Pearson r (7∼8B)**	**Pearson r (All 17)**
R-Avg (`V`) vs. FR (`P`)	0.836	0.832
R-Avg (`V`) vs. NFR (`I`)	0.824	0.774
R-Avg (`V`) vs. NFR (`D`)	0.805	0.764
FR (`P`) vs. NFR (`I`)	0.748	0.715

5 Anti-Distillation Policy Optimization (ADPO)

LMMs faithfully perceive video content but systematically underuse it during causal discovery, defaulting to textual priors. Mainstream rule-based RL methods like Group Relative Policy Optimization (GRPO) [11] struggle to address this, as the verifier only checks answers without distinguishing whether correctness arises from visual understanding or textual co-occurrence, while KL constraint anchors the policy to a biased initialization. We propose Anti-Distillation Policy Optimization (ADPO), which introduces a negative teacher alignment paradigm by explicitly repelling the policy from a prior-only counterfactual teacher, forcing the reasoning distribution to ground in visual evidence.

5.1 Preliminaries: Group Relative Policy Optimization (GRPO)

Given a video-question pair (V, q), the policy π_θ samples a group of G rollouts $\{o_i\}_{i=1}^{G}$, each scored by a rule-based verifier to obtain a scalar reward r_i. The group-relative advantage is computed as $A_i = (r_i - \text{mean}(\{r\}))/\text{std}(\{r\})$, and the policy maximizes a clipped surrogate objective regularized by KL divergence toward a reference model π_{ref}:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\Big[\min(\rho_{i,t}A_i,\ \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)A_i) - \beta\,\mathbb{D}_{\text{KL}}[\pi_\theta\|\pi_{\text{ref}}]\Big]\right], \quad (5)$$

where $\rho_{i,t} = \pi_\theta(o_{i,t} \mid V, q, o_{i,<t})/\pi_{\theta_{\text{old}}}(o_{i,t} \mid V, q, o_{i,<t})$ is the importance sampling ratio.

5.2 Framework of Anti-Distillation Policy Optimization

ADPO comprises three stages, namely visual perturbation, perception-sensitivity regularization, and joint optimization. Specifically, we apply perturbations to the input video to construct counterfactual views, evaluate the policy on both the original and perturbed inputs respectively, and penalize outputs that are unresponsive to visual changes, thereby suppressing prior-driven shortcut behaviors.

Table 4: Experimental results of ADPO compared with baselines across different backbones. We additionally report three variants corresponding to different visual perturbation operators $\mathcal{P}$.

Backbone	Method	Veridicality	Plausibility		Incongruence			Deprivation		
		R-Avg ↑	R-Avg ↑	FR ↓	R-Avg ↓	Acc ↑	NFR ↓	R-Avg ↓	Acc ↑	NFR ↓
Qwen3-VL-4B	Base [2]	30.2	1.7	33.3	29.7	11.4	50.4	26.5	35.0	36.7
	GRPO [11]	32.8	2.9	39.5	38.8	8.6	59.7	28.7	29.6	41.0
	DAPO [31]	34.1	1.8	43.5	33.1	12.0	53.6	21.6	41.2	27.5
	ADPO (Ours)	**36.4**	**3.4**	**25.7**	**28.1**	**14.0**	**39.0**	**17.5**	**47.4**	**26.0**
Qwen3-VL-8B	Base [2]	36.4	3.9	51.7	31.2	13.2	54.1	36.7	24.6	52.5
	GRPO [11]	41.3	4.2	52.1	38.9	10.0	60.6	38.4	20.8	55.9
	DAPO [31]	39.2	2.6	48.9	39.0	9.4	54.3	37.9	20.8	58.3
	ADPO w/ *Shuffle* (Ours)	**43.1**	**4.2**	**42.2**	**30.9**	**17.2**	**47.0**	**33.7**	**32.0**	**43.1**
	ADPO w/ *Crop*	41.9	3.9	46.3	33.9	15.0	53.9	37.1	25.0	47.3
	ADPO w/ *Replace*	41.4	4.0	45.5	31.6	14.4	51.2	36.6	27.0	49.3

5.2.1 Visual Perturbation

For every training instance (V, q), we construct a corrupted counterpart $\tilde{V}$ by applying a stochastic perturbation operator $\mathcal{P}(\cdot)$ to the original video. Specifically, $\mathcal{P}$ is selected from a set of three visual perturbations, namely frame shuffling, spatial cropping, and frame replacement (see Appendix C for implementation details), each disrupting a distinct aspect of the visual evidence:

$$\tilde{V} \;=\; \mathcal{P}(V), \quad \mathcal{P} \in \{\mathcal{P}_{\text{shuffle}}, \mathcal{P}_{\text{crop}}, \mathcal{P}_{\text{replace}}\}, \tag{6}$$

where $\mathcal{P}_{\text{shuffle}}$ randomly permutes the temporal order of a fixed number of frames, $\mathcal{P}_{\text{crop}}$ randomly masks a fixed-area region within each frame, and $\mathcal{P}_{\text{replace}}$ replaces a fixed number of randomly selected frames with blank frames. The choice of $\mathcal{P}$ is fixed throughout a single training run as a configuration hyperparameter, while its internal randomness is freshly resampled at every optimization step. The corrupted video retains enough low-level statistics to remain in-distribution for the vision encoder, while partially disrupting the fine-grained visual evidence that reasoning would otherwise rely upon.

5.2.2 Perception-Sensitivity Regularization

At each iteration, ADPO performs two forward passes to recompute the distributions over the same group of rollouts $\{o_i\}_{i=1}^{G}$ generated from the original video, one conditioned on the original visual input V and the other on its perturbed counterpart $\tilde{V}$, yielding two policy distributions respectively:

$$\pi_\theta^{(i,t)} = \pi_\theta(\cdot \mid V, q, o_{i,<t}), \quad \pi_\theta^{\text{pert},(i,t)} = \pi_\theta(\cdot \mid \tilde{V}, q, o_{i,<t}). \tag{7}$$

This ensures that any divergence between $\pi_\theta^{(i,t)}$ and $\pi_\theta^{\text{pert},(i,t)}$ is unambiguously attributable to the model's sensitivity to the visual modality, rather than being confounded by differences in the trajectory. We then define the perception-sensitivity regularizer as a repulsion objective against the prior-only counterfactual teacher:

$$\mathcal{L}_{\text{prcp}}(\theta) \;=\; \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\mathbb{D}_{\text{KL}}\Big[\,\pi_\theta^{(i,t)} \,\|\, \text{sg}\left(\pi_\theta^{\text{pert},(i,t)}\right)\Big], \tag{8}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator, which enforces asymmetric repulsion from the prior-only counterfactual teacher $\pi_\theta^{\text{pert},(i,t)}$, forcing the grounded distribution $\pi_\theta^{(i,t)}$ to move away from the prior-dominated policy rather than allowing both distributions to collapse toward arbitrary midpoint.

5.2.3 Joint Optimization

Combining the standard GRPO surrogate with the perception-sensitivity term, we maximize:

$$\mathcal{J}_{\text{ADPO}}(\theta) \;=\; \mathcal{J}_{\text{GRPO}}(\theta) \;+\; \lambda\,\mathcal{L}_{\text{prcp}}(\theta), \tag{9}$$

where λ is a single hyperparameter controlling the strength of the anti-distillation regularizer. The detailed hyperparameter settings analysis are provided in Appendix C.

6 Experiments

6.1 Experimental Setups

We compare our ADPO against two RLVR baselines, namely GRPO [11] and DAPO [31]. Note that GRPO employs a reference KL penalty, while DAPO removes it and adopts dynamic sampling instead.

We evaluate the effectiveness of the baselines and our ADPO on two LMMs, namely Qwen3-VL-4B and 8B [2]. See Appendix C for more implementation details.

6.2 Experiment Result

Table 4 reports the performance of ADPO against competing RLVR baselines. ADPO not only attains the strongest overall performance but also exhibits a property absent in both GRPO and DAPO, concurrently enhancing causal discovery accuracy and robustness to perturbation. While GRPO and DAPO improve `Veridicality` performance, such gains are accompanied by an intensified dependence on textual priors, as evidenced by elevated FR under `Plausibility` and higher NFR under `Incongruence` and `Deprivation`. In contrast, ADPO reverses this trade-off, markedly suppressing FR while attaining the highest Acc and lowest NFR across perturbed configurations. These findings suggest that conventional RLVR methods achieve apparent accuracy gains by amplifying prior reliance, whereas ADPO explicitly targets and mitigates this failure mode, yielding more reliable multimodal causal discovery grounded in visual evidence.

6.3 Analysis

Complementary Evidence of Visual Engagement. Figure 4 examines the internal behavior of different strategies from two complementary perspectives. As indirect evidence at the output level, ADPO widens the likelihood gap between correct and shortcut answers, indicating that our method endows the model with a stronger discriminative capability against prior-driven cues. As direct evidence at the attention level, ADPO allocates markedly higher weight mass to video tokens during chain-of-thought, demonstrating that it actively restructures the model's internal information flow and channels reasoning more through visual evidence rather than textual priors.

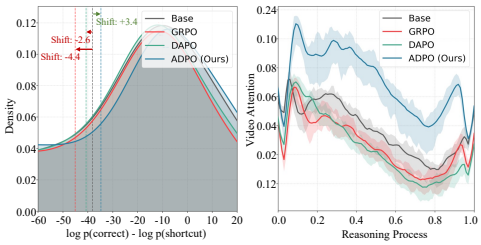


Figure 4: Discriminative capability between correct and shortcut answers measured by log-likelihood gap (left) and attention weights on visual tokens under different methods (right).

Effect of Different Visual Perturbation Strategies. As shown in Table 4, all three perturbation strategies consistently outperform the baseline across the four evaluation configurations. Among them, frame shuffling yields the most significant improvement, as it disrupts the temporal causal order between events, providing the sharpest contrastive signal for the perception-sensitivity regularizer. These results indicate that temporal dynamics constitute a critical dimension of video information, and disrupting the temporal order is the most effective way to corrupt video evidence.

General Video Understanding Capability. A natural concern is whether the perception-sensitivity regularizer, while improving causal discovery, compromises the model's general video understanding capability. To address this, we evaluate ADPO on a suite of standard video benchmarks, namely Video-MME [32], LongVideoBench [33], MMVU [34], and MVBench [35], covering perception and reasoning over different video durations. As shown in Table 5, ADPO achieves better across all benchmarks, indicating that the anti-distillation objective does not come at the cost of fundamental comprehension.

Table 5: General video understanding performance.

Method	Video-MME	LongVideoBench	MMVU	MVBench
Qwen3-VL-8B	64.7	56.3	69.7	65.3
w/ GRPO	65.4	60.5	70.2	66.0
w/ DAPO	65.9	60.1	72.0	67.1
w/ **ADPO** (Ours)	**66.7**	**62.3**	**73.1**	**68.7**

7 Conclusion

In this paper, we shift the evaluation of causal discovery in Large Multimodal Models (LMMs) from outcome accuracy to mechanism diagnosis through ProCauEval, a perturbation-based protocol that decomposes each modality's contribution via five controlled configurations. Evaluating 17 mainstream LMMs, we find that visual comprehension is preserved while causal reasoning bypasses it for textual priors, and this deficit amplified by reasoning-oriented post-training and most pronounced in top-performing models. Building on this diagnosis, we propose Anti-Distillation Policy Optimization (ADPO), which introduce the prior-only distribution as a negative teacher augments GRPO with

a perception-sensitivity regularizer that maximizes divergence between policy distributions under original and visually corrupted inputs. Extensive experiments demonstrate that ADPO grounds reasoning in visual evidence, offering a preliminary step toward reliable multimodal causal discovery.

Acknowledgments

The research in this article is supported by the National Science Foundation of China (U22B2059, 62276083), Key Research and Development Program of Heilongjiang Province (2024ZX01A05) , Suzhou Municipal Key Technology R&D Project / Suzhou Municipal Key Core Technology Project (SYG2025069).

References

[1] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=zKv8qULV6n.

[2] 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*, 2025.

[3] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[4] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. URL http://arxiv.org/abs/1705.06950.

[5] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9777–9786. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00965. URL https://openaccess.thecvf.com/content/CVPR2021/html/Xiao_NExT-QA_Next_Phase_of_Question-Answering_to_Explaining_Temporal_Actions_CVPR_2021_paper.html.

[6] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5288–5296. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.571. URL https://doi.org/10.1109/CVPR.2016.571.

[7] Tieyuan Chen, Huabin Liu, Yi Wang, Yihang Chen, Tianyao He, Chaofan Gan, Huanyu He, and Weiyao Lin. Mecd+: Unlocking event-level causal graph discovery for video reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[8] Xuchen Li, Xuzhao Li, Shiyu Hu, Kaiqi Huang, and Wentao Zhang. Causalstep: A benchmark for explicit stepwise causal reasoning in videos. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor, editors, *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 6530–6538. AAAI Press, 2026. doi: 10.1609/AAAI.V40I8.37582. URL https://doi.org/10.1609/aaai.v40i8.37582.

[9] Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T. Kao. Causalvqa: A physically grounded causal reasoning benchmark for video models. *CoRR*, abs/2506.09943, 2025. doi: 10.48550/ARXIV.2506.09943. URL https://doi.org/10.48550/arXiv.2506.09943.

[10] Pritam Sarkar and Ali Etemad. Vcrbench: Exploring long-form causal reasoning capabilities of large video language models. *CoRR*, abs/2505.08455, 2025. doi: 10.48550/ARXIV.2505.08455. URL https://doi.org/10.48550/arXiv.2505.08455.

[11] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948. URL https://doi.org/10.48550/arXiv.2501.12948.

[12] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL http://arxiv.org/abs/1503.02531.

[13] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 21241–21250. IEEE, 2022. doi: 10.1109/CVPR52688.2022.02059. URL https://doi.org/10.1109/CVPR52688.2022.02059.

[14] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can I trust your answer? visually grounded video question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13204–13214. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01254. URL https://doi.org/10.1109/CVPR52733.2024.01254.

[15] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *CoRR*, abs/2503.06749, 2025. doi: 10.48550/ARXIV.2503.06749. URL https://doi.org/10.48550/arXiv.2503.06749.

[16] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-VL: learning to reason with multimodal large language models via step-wise group relative policy optimization. *CoRR*, abs/2503.12937, 2025. doi: 10.48550/ARXIV.2503.12937. URL https://doi.org/10.48550/arXiv.2503.12937.

[17] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *CoRR*, abs/2503.10615, 2025. doi: 10.48550/ARXIV.2503.10615. URL https://doi.org/10.48550/arXiv.2503.10615.

[18] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *CoRR*, abs/2503.06749, 2025. doi: 10.48550/ARXIV.2503.06749. URL https://doi.org/10.48550/arXiv.2503.06749.

[19] Yi-Fan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, et al. R1-reward: Training multimodal reward model through stable reinforcement learning. *arXiv preprint arXiv:2505.02835*, 2025.

[20] Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, et al. Scaling rl to long videos. *arXiv preprint arXiv:2507.07966*, 2025.

[21] Team Glm, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

[22] Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, et al. Onethinker: All-in-one reasoning model for image and video. *arXiv preprint arXiv:2512.03043*, 2025.

[23] Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025.

[24] 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. *arXiv preprint arXiv:2503.21776*, 2025.

[25] Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *arXiv preprint arXiv:2509.21100*, 2025.

[26] Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Xiao-Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, Zihan Zhang, Zihan Jiang, Zhixian Zheng, Zhichao Song, Zhenbo Luo, Yue Yu, Yudong Wang, Yuanyuan Tian, Yu Tu, Yihan Yan, Yi Huang, Xu Wang, Xinzhe Xu, Xingchen Song, Xing Zhang, Xing Yong, Xin Zhang, Xiangwei Deng, Wenyu Yang, Wenhan Ma, Weiwei Lv, Weiji Zhuang, Wei Liu, Sirui Deng, Shuo Liu, Shimao Chen, Shihua Yu, Shaohui Liu, Shande Wang, Rui Ma, Qiantong Wang, Peng Wang, Nuo Chen, Menghang Zhu, Kangyang Zhou, Kang Zhou, Kai Fang, Jun Shi, Jinhao Dong, Jiebao Xiao, Jiaming Xu, Huaqiu Liu, Hongshen Xu, Heng Qu, Haochen Zhao, Hanglong Lv, Guoan Wang, Duo Zhang, Dong Zhang, Di Zhang, Chong Ma, Chang Liu, Can Cai, and Bingquan Xia. Mimo-vl technical report. *CoRR*, abs/2506.03569, 2025. doi: 10.48550/ARXIV.2506.03569. URL https://doi.org/10.48550/arXiv.2506.03569.

[27] OpenAI. Openai GPT-5 system card. *CoRR*, abs/2601.03267, 2026. doi: 10.48550/ARXIV.2601.03267. URL https://doi.org/10.48550/arXiv.2601.03267.

[28] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

[30] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507, 2025.

[31] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: an open-source LLM reinforcement learning system at scale. *CoRR*, abs/2503.14476, 2025. doi: 10.48550/ARXIV.2503.14476. URL https://doi.org/10.48550/arXiv.2503.14476.

[32] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24108–24118. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02245. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

[33] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In Amir Globersons, Lester Mackey,

Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/329ad516cf7a6ac306f29882e9c77558-Abstract-Datasets_and_Benchmarks_Track.html.

[34] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 8475–8489. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00793. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhao_MMVU_Measuring_Expert-Level_Multi-Discipline_Video_Understanding_CVPR_2025_paper.html.

[35] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22195–22206. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02095. URL https://doi.org/10.1109/CVPR52733.2024.02095.

[36] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. *CoRR*, abs/2407.10671, 2024. doi: 10.48550/ARXIV.2407.10671. URL https://doi.org/10.48550/arXiv.2407.10671.

[37] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01100. URL https://doi.org/10.1109/ICCV51070.2023.01100.

[38] Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

[39] Michael Tschannen, Alexey A. Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier J. Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *CoRR*, abs/2502.14786, 2025. doi: 10.48550/ARXIV.2502.14786. URL https://doi.org/10.48550/arXiv.2502.14786.

[40] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan 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. *CoRR*, abs/2502.13923, 2025. doi: 10.48550/ARXIV.2502.13923. URL https://doi.org/10.48550/arXiv.2502.13923.

[41] Bingquan Xia, Bowen Shen, Cici, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaqiu Liu, Jiebao Xiao, Jinhao Dong, Liang Zhao, Peidian Li, Peng Wang, Shihua Yu, Shimao Chen, Weikun Wang, Wenhan Ma, Xiangwei Deng, Yi Huang, Yifan Song, Zihan Jiang, Bowen Ye,

Can Cai, Chenhong He, Dong Zhang, Duo Zhang, Guoan Wang, Hao Tian, Haochen Zhao, Heng Qu, Hongshen Xu, Jun Shi, Kainan Bao, QingKai Fang, Kang Zhou, Kangyang Zhou, Lei Li, Menghang Zhu, Nuo Chen, Qiantong Wang, Shaohui Liu, Shicheng Li, Shuhao Gu, Shuhuai Ren, Shuo Liu, Sirui Deng, Weiji Zhuang, Weiwei Lv, Wenyu Yang, Xin Zhang, Xing Yong, Xing Zhang, Xingchen Song, Xinzhe Xu, Xu Wang, Yihan Yan, Yu Tu, Yuanyuan Tian, Yudong Wang, Yue Yu, Zhenru Lin, Zhichao Song, and Zihao Yue. Mimo: Unlocking the reasoning potential of language model - from pretraining to posttraining. *CoRR*, abs/2505.07608, 2025. doi: 10.48550/ARXIV.2505.07608. URL https://doi.org/10.48550/arXiv.2505.07608.

[42] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Yihui He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=wCXAlfvCy6.

[43] Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from GLM-130B to GLM-4 all tools. *CoRR*, abs/2406.12793, 2024. doi: 10.48550/ARXIV.2406.12793. URL https://doi.org/10.48550/arXiv.2406.12793.

[44] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor G. Turrisi da Costa, Louis Béthune, Zhe Gan, Alexander Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua M. Susskind, and Alaaeldin El-Nouby. Multimodal autoregressive pre-training of large vision encoders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 9641–9654. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00901. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fini_Multimodal_Autoregressive_Pre-training_of_Large_Vision_Encoders_CVPR_2025_paper.html.

[45] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[46] Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

A Further Discussion, Limitations, and Future Work

While our work introduces PROCAUEVAL and ADPO with promising results, we acknowledge several limitations that suggest avenues for future work:

Language coverage. Our evaluation is conducted exclusively in English. Although the proposed protocol is language-agnostic in principle, extending PROCAUEVAL to multilingual settings, particularly low-resource languages, remains an interesting direction for future investigation.

Video duration distribution. The videos in our benchmark range from 8.9s to 536.9s, with an average duration of 132.8s. While this range covers a reasonable spectrum of typical video content, ultra-long videos (e.g., feature-length films exceeding one hour) are not represented in our current evaluation set. Whether the observed causal discovery deficit persists at substantially longer temporal scales remains to be verified.

Frame sampling rate during training. Due to computational constraints, we uniformly sample 16 frames per video during ADPO training. Although we use 128 frames at evaluation time to ensure fair comparison, exploring denser frame sampling during training could potentially yield further improvements.

Evaluation cost of GPT Score. The `Diagnosis` configuration relies on GPT-based scoring for caption quality assessment. While this provides a fine-grained evaluation, it incurs API costs and introduces a dependency on a closed-source model. Future work could explore open-source alternatives for caption quality assessment.

B Evaluation Settings

B.1 Fake Events Construction and Potential Leakage Concern

Fake events are designed as controlled textual perturbations for diagnosing whether the model genuinely relies on visual evidence during causal discovery. Rather than serving as ordinary distractors, they are constructed to be plausible shortcut candidates under a prior-only inference setting. Specifically, we retain fabricated causes that receive higher language-only log-probability than the original causes, ensuring that the model would be likely to select them if visual evidence were not effectively used. This allows `Plausibility` to distinguish reliable visual grounding from prior-driven reasoning: If the shortcut is already attractive under text-only inference, resisting it under multimodal input provides stronger evidence that the model is reasoning over visual content.

B.2 Fairness of Model-Specific Fake Event Selection

A potential concern is that selecting fake events based on each model's own log-probabilities could yield different perturbation sets across models, raising questions about cross-model comparability. To eliminate this concern, we adopt a unified fake event set constructed as the intersection of model-preferred candidates across all evaluated LMMs. Specifically, for each video, we retain only those fabricated cause events that receive higher language-only log-probability than the ground-truth causes under every model in our evaluation pool, ensuring that all models face an identical set of perturbations. This intersection-based design guarantees strict cross-model comparability: Every model is evaluated against the same pool of fake events, eliminating confound introduced by model-specific selection. Moreover, an event surviving this intersection is, by construction, a statistically tempting shortcut for all models simultaneously, which makes it a stronger and more universal probe of textual prior reliance than any per-model selection. Consequently, the FR metric reported in our evaluation reflects a uniform and rigorous measurement of shortcut dependence that is directly comparable across the entire model suite.

B.3 Quality Validation of Fabricated Video Clips

We validate each synthesized clip along two target dimensions using CLIP embeddings [29]. For semantic relevance, we compute cross-modal cosine similarity between each clip and its corresponding event description, regenerating clips below 0.30 until the threshold is met. For scene inconsistency, we compute intra-video visual similarity between each synthesized clip and the remaining real clips, regenerating clips above 0.70. This iterative generation enforces both intended properties, yielding clips semantically meaningful yet visually distinguishable.

B.4 Detailed Descriptions of LMMs

LLaVA-OneVision [1] is a model consolidating insights from the LLaVA-NeXT series, built on Qwen-2 [36] and SigLIP [37] vision encoder. It employs a Higher AnyRes strategy with balanced token allocation to flexibly represent visual signals across single-image, multi-image, and video scenarios. The model follows a curriculum learning process consisting of language-image alignment, high-quality knowledge learning, and a OneVision stage blending 1.6M mixed-modality samples, enabling strong task transfer from images to videos.

Qwen3-VL [2] is built on Qwen3 [38] and SigLIP-2 encoder [39], supporting 256K-token interleaved contexts. With Interleaved MROPE, DeepStack, and text timestamps, it achieves leading results on reasoning benchmarks.

InternVL3.5 [3] is built on Qwen3 and InternViT vision encoder, proposing Cascade RL that combines offline and online reinforcement learning to enhance reasoning. With ViR dynamically adjusting visual token resolution and DvD decoupling vision-language deployment, it achieves 4.05× inference speedup.

MiMo-VL-SFT [26] is built on Qwen2.5-ViT [40] and MiMo-7B-Base [41] (Xiaomi, 2025), undergoing four-stage pre-training with 2.4T tokens. Incorporating long Chain-of-Thought reasoning data into late pre-training stages.

GPT-5.1 [27] is OpenAI's unified multimodal system featuring a fast main model and a deep reasoning model with real-time routing. Supporting text, image inputs, and tool use, it employs safe-completions safety training. It significantly outperforms predecessors on HealthBench with 26-65% hallucination reduction.

Video-R1 [24] is built on Qwen2.5-VL-7B, pioneering the extension of R1 reinforcement learning to video reasoning. It proposes T-GRPO to encourage temporal modeling by comparing ordered and shuffled frame sequences.

VideoChat-R1.5 [25] is built on Qwen2.5-VL, proposing Visual Test-Time Scaling that refines spatio-temporal focus through iterative perception. Using GRPO-based reinforcement learning to optimize perception policies, it achieves over 5% average improvement across 15+ benchmarks on video QA and spatio-temporal perception.

Keye-VL-1.5 [25] is built on Qwen3-8B and SigLIP, proposing Slow-Fast video encoding that dynamically allocates resolution based on inter-frame similarity. Through four-stage progressive pre-training extending context from 8K to 128K, it outperforms most comparable models.

LongVILA-R1 [20] is built on LongVILA [42], pioneering the extension of reinforcement learning to long video reasoning. With a 104K long video QA dataset and two-stage training (CoT-SFT + GRPO), and MR-SP system supporting 8,192 frames, it achieves 65.1% on VideoMME [32].

GLM-4.1V-Thinking [30] is built on GLM-4-9B [43] and AIMv2-Huge vision encoder [44], proposing RLCS to enhance multimodal reasoning. With cross-domain joint training across STEM, video, and GUI agents, it outperforms Qwen2.5-VL-72B on 29 benchmarks while maintaining 9B scale.

Qwen3-VL-Thinking-8B [2] is built on Qwen3 and SigLIP-2 encoder, enhancing reasoning through RL following long CoT cold-start. With Interleaved MROPE and DeepStack architecture, it significantly outperforms the base instruct version on reasoning benchmarks like MathVista [45].

OneThinker [22] is built on Qwen3-VL-Instruct-8B, unifying 10 visual tasks including QA, captioning, spatio-temporal grounding, tracking, and segmentation. With OneThinker-600k corpus and EMA-GRPO addressing reward heterogeneity in multi-task RL, it achieves strong performance across 31 benchmarks while enabling cross-task knowledge transfer.

MiMo-VL-RL [26] is built on Qwen2.5-ViT and MiMo-7B-Base, further optimized from SFT via Mixed On-policy Reinforcement Learning (MORL). With joint RLVR and RLHF training integrating diverse reward signals, it achieves 66.7% on MMMU and 59.4% on OlympiadBench.

GPT-5.1-Thinking [27] is trained via reinforcement learning to produce long internal chains of thought before answering, refining its reasoning and recognizing mistakes. Supporting parallel test-time compute with browsing, coding, and tool use, it achieves 46.2% on HealthBench Hard [46].

C Anti-Distillation Policy Optimization

C.1 Implementation Details

All experiments are conducted on 4 NVIDIA A100 GPUs with a global batch size of 64 with learning rate of $1e{-}6$. Models are trained for 1 epoch on 20K QA pairs sampled from Video-R1-260k [24]. Videos are uniformly sampled to 16 frames during training and 128 frames during evaluation. The group size for rollout is set to $G = 5$, and the anti-distillation coefficient is set to $\lambda = 0.01$. All video corruption ratios are set to 0.2. All other hyperparameters follow the default GRPO. The reported results are averaged over 3 independent runs with different random seeds.

Prompt for GPT Score

<Video> **Video** </Video >

You are an expert evaluator for video captions. Given a video description (caption), evaluate its quality across the following dimensions. Score each dimension from 1 to 10.

Video Caption: {caption}

Evaluate the caption on these dimensions:

1. **Accuracy (1-10):** Does the caption correctly describe the visual content without hallucination?
   - 1-2: Major factual errors or hallucinations throughout
   - 3-4: Several inaccuracies or noticeable hallucinations
   - 5-6: Mostly correct with minor inaccuracies
   - 7-8: Accurate with only trivial errors
   - 9-10: Fully accurate, no errors or hallucinations

2. **Completeness (1-10):** Does the caption cover the key events, actions, and objects in the video?
   - 1-2: Misses most important content
   - 3-4: Covers only a small portion of key content
   - 5-6: Covers main events but misses notable details
   - 7-8: Covers most content with only minor omissions
   - 9-10: Comprehensively covers all important content

Think step by step before scoring. Provide your evaluation in the following JSON format:

```
{
  "accuracy": <1-10>,
  "completeness": <1-10>,
  "overall": <weighted average, 1-10>,
  "reasoning" : <brief explanation covering each dimension>
}
```

Figure 5: Prompt for GPT Score

C.2 Gradient Analysis

A natural concern with the joint objective in Eq. 9 is whether the GRPO advantage-weighted gradient and the perception-sensitivity regularizer may push the policy in conflicting directions. We provide a gradient-level analysis to characterize their interaction and identify when conflicts may arise.

Gradient Decomposition. Taking the gradient of $J_{\text{ADPO}}(\theta)$ with respect to θ yields:

$$\nabla_\theta J_{\text{ADPO}} = \underbrace{\mathbb{E}\left[\sum_{i,t} A_i \, \nabla_\theta \log \pi_\theta(o_{i,t} \mid V, q, o_{i,<t})\right]}_{\nabla_\theta J_{\text{GRPO}}} + \lambda \underbrace{\nabla_\theta \mathbb{E}\left[D_{\text{KL}}\left(\pi_\theta^{(i,t)} \,\|\, \text{sg}(\pi_\theta^{\text{pert},(i,t)})\right)\right]}_{\nabla_\theta \mathcal{L}_{\text{prcp}}}. \tag{10}$$

The stop-gradient operator $\text{sg}(\cdot)$ ensures that the regularizer's gradient flows only through the visually-grounded distribution $\pi_\theta(\cdot \mid V, q, o_{i,<t})$, explicitly pushing it away from the prior-dominated counterpart $\pi_\theta(\cdot \mid \tilde{V}, q, o_{i,<t})$.

Case-wise Interaction Analysis. Whether the two gradient terms cooperate or conflict depends on the source of the rollout's correctness. We enumerate four canonical cases in Table 6, characterized by (i) whether the rollout is supported by visual evidence or driven by textual priors, and (ii) the sign of the advantage A_i.

Table 6: Interaction between GRPO and perception-sensitivity gradients across rollout types.

Case	Rollout source	A_i	GRPO direction	$\mathcal{L}_{\text{prcp}}$ direction
(a)	Visually-grounded, correct	$+$	Increase $\pi_\theta(o \mid V)$	Push away from prior
(b)	Visually-grounded, incorrect	$-$	Decrease $\pi_\theta(o \mid V)$	Push away from prior
(c)	Prior-driven, correct (coincidental)	$+$	Increase $\pi_\theta(o \mid V)$	Push away from prior
(d)	Prior-driven, incorrect	$-$	Decrease $\pi_\theta(o \mid V)$	Push away from prior

In cases (a), (b), and (d), the two gradients are mutually compatible, the regularizer reinforces the GRPO signal by either amplifying visually-grounded behaviors or suppressing prior-driven ones. The only potentially conflicting case is (c), where a rollout happens to be correct due to alignment

between textual priors and the ground truth, and the GRPO term reinforces this rollout while $\mathcal{L}_{\text{prcp}}$ pushes the distribution away from the prior. We argue that this conflict is benign for the reason of distinct gradient subspaces: The GRPO gradient operates on the log-probability of a specific token $o_{i,t}$, increasing its probability mass locally. In contrast, $\mathcal{L}_{\text{prcp}}$ operates on the shape of the full distribution $\pi_\theta(\cdot \mid V, q, o_{i,<t})$, redistributing mass across the vocabulary. The two gradients therefore live in largely orthogonal subspaces, $\mathcal{L}_{\text{prcp}}$ does not directly suppress the probability of the specific correct token in case (c), but rather reshapes the distribution to differ from the prior elsewhere, which is compatible with concentrating mass on visually-supported tokens.

C.3 Sensitivity to the Anti-Distillation Coefficient λ

A central claim of ADPO is that it breaks the accuracy-robustness trade-off observed in GRPO and DAPO, where higher `Veridicality` R-Avg is consistently accompanied by elevated FR and NFR (shown in Table 2). To verify this claim, we conduct a sensitivity analysis over the anti-distillation coefficient λ. We train Qwen3-VL-8B with frame shuffling perturbation under $\lambda \in \{0, 0.001, 0.005, 0.01, 0.02, 0.03, 0.04\}$, where $\lambda = 0$ recovers vanilla GRPO. All other hyperparameters are kept identical to the main experiments.

Table 7: Sensitivity of ADPO to the anti-distillation coefficient λ on Qwen3-VL-8B. $\lambda = 0$ corresponds to vanilla GRPO. The shaded region marks the regime where both `Veridicality` accuracy and perturbation robustness are simultaneously improved over GRPO.

λ	`Veridicality`	`Plausibility`	`Incongruence`	`Deprivation`	Regime
	R-Avg ↑	FR ↓	NFR ↓	NFR ↓	
0 (GRPO)	41.3	52.1	60.6	55.9	Baseline
0.001	41.7	50.4	58.3	54.1	Under-reg.
0.005	42.6	44.8	50.1	46.3	Sweet spot
0.01	**43.1**	**42.2**	**47.0**	**43.1**	Sweet spot
0.02	42.4	43.5	48.6	44.8	Sweet spot
0.03	40.8	46.7	51.9	48.7	Over-reg.
0.04	36.2	50.3	56.4	53.2	Collapse

Three Regimes of λ. The results in Table 7 reveal three distinct regimes:

- **(i) Under-regularization ($\lambda \leq 0.001$).** The regularizer is too weak to meaningfully reshape the policy, and ADPO's behavior closely tracks GRPO. The trade-off between `Veridicality` accuracy and perturbation robustness remains visible.
- **(ii) Sweet spot ($\lambda \in [0.005, 0.02]$).** Within this range, ADPO simultaneously improves `Veridicality` R-Avg over GRPO and reduces FR / NFR. The trade-off between accuracy and robustness is genuinely broken, validating our central claim.
- **(iii) Over-regularization ($\lambda \geq 0.03$).** The regularizer begins to dominate the GRPO objective, indiscriminately pushing the policy away from any prior-aligned behavior, including the case (c) rollouts discussed in Section C.2. This causes `Veridicality` accuracy to degrade below the GRPO baseline, while robustness gains also erode because the policy becomes unstable. At $\lambda = 0.04$, performance collapses toward or below baseline across all configurations.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper's contributions and scope can be found in the section of Abstract, Section 1, and Section 2.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work are described in Appendix A

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not contain formal theoretical results, theorems, or proofs. Our contributions are empirical and methodological, consisting of a perturbation-based evaluation protocol (PROCAUEVAL) and a reinforcement learning method (ADPO).

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Implementation Details can be found in Section 3, Section 5, and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper builds upon publicly available datasets (MECD, Video-R1-260k) and open-source models (Qwen3-VL, etc.), but the current submission does not provide an anonymized code release and data. The evaluation protocol (Section 2 and Appendix B), the ADPO algorithm (Section 4), and the experimental setup (Appendix C.1) are described in detail. Code, data, and model checkpoints will be released upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Implementation Details can be found in Section 3, Section 5, and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Statistical significance of the experiments can be found in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Implementation Details can be found in Section 3, Section 5, and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We briefly discuss the broader impacts of our work here. On the positive side, ProCauEval provides the community with a diagnostic tool that exposes the gap between apparent and genuine causal reasoning in LMMs, which is particularly valuable for safety-critical applications such as autonomous driving and medical video analysis where

reliance on textual priors rather than visual evidence could lead to incorrect causal judgments. ADPO further offers a principled training objective that grounds model reasoning in visual evidence, reducing the risk of overconfident yet visually unsupported predictions. On the negative side, our diagnostic protocol explicitly identifies failure modes (e.g., susceptibility to fabricated cause events) that could in principle be exploited by adversaries to construct misleading inputs for deployed systems. However, we believe that openly characterizing these failure modes is a necessary step toward developing more robust models, and the benefits of public awareness outweigh the risks of obscurity. The fake video clips generated by VEO-3 in our protocol are restricted to research use and are not intended for deployment in any generative content pipeline.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our released artifacts do not pose a high risk of misuse. PROCAUEVAL is a diagnostic protocol, and the fabricated video clips generated by VEO-3 are short, and deliberately scene-inconsistent, making them unsuitable as a general-purpose video generation resource. ADPO is a training objective applied on top of open-source models and does not introduce additional generative capabilities.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The creators or original owners of assets (e.g., code, data, models), used in the paper, are properly credited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: The new assets (PROCAUEVAL data and ADPO code) are not released at submission time. The paper documents the data construction pipeline (Section 3 and Appendix B), the algorithm (Section 5), and training details (Appendix C.1). Full documentation will be provided alongside the assets upon release.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM is used only for writing, editing, or formatting purposes.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.