LAMP: LEARNABILITY- AND HORIZON-AWARE ON-POLICY DISTILLATION FROM PRIVILEGE MODALITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) often struggle with perceptual information that is difficult to recover from RGB alone. Cross-modal on-policy distillation (OPD) addresses this issue by allowing the teacher to access privilege modalities such as depth or video, but existing methods treat token-level supervision uniformly, conflating transferable reasoning with unlearnable sensor information. Therefore, we propose LAMP, a learnability- and horizon-aware privilege-modality on-policy distillation framework for MLLMs. LAMP makes a two-fold contribution: First, it performs counterfactual modality attribution to identify whether each token's teacher advantage is grounded in the student or privilege modality, enabling a self-calibrating gate to suppress untransferable supervision. Second, it introduces horizon-conditioned divergence weighting to emphasize tokens that have greater downstream influence on the reasoning trajectory. Extensive experiments on depth-to-RGB and video-to-RGB transfer show that LAMP consistently outperforms standard OPD and representative cross-modal distillation methods, achieving up to 6.1% and 10.7% performance gain over the strongest baselines on depth and video benchmarks, respectively.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have achieved remarkable success in visual understanding and vision-language reasoning (Yan et al., 2025). Yet this success masks a systematic limitation: when tasks demand precise perception, such as estimating depth, tracking motion, MLLMs frequently fail (Kancheti et al., 2026). One difficulty is that some task-relevant information is encoded implicitly in an RGB image and requires stronger understanding and reasoning capabilities. Post-training offers a useful framework for improving such capabilities, and on-policy distillation (OPD) has emerged as a promising paradigm (Agarwal et al., 2024; Xu et al., 2025; Jin et al., 2026; Kim & Baek, 2026). OPD trains the student on its own trajectories while providing dense token-level supervision from a teacher, making it well suited for improving the student's reasoning process. Its effectiveness has also led to adoption in recent frontier open-source models, including DeepSeek-V4 (DeepSeek-AI et al., 2026), GLM-5 (GLM-5 Team et al., 2026), and Kimi K3 (Kimi Team et al., 2026). However, standard OPD does not fully fit our setting, where the information needed for supervision may be unavailable in the student's modality.

Targeting this setting, cross-modal OPD offers an effective solution. As shown in Figure 1, the teacher accesses a modality the student does not possess, providing the implicit information explicitly during training while inference remains RGB image only. Recent work has begun to explore similar cross-modal scenario. X³-OPD (Fu et al., 2026) transfers reasoning from a powerful text teacher to an audio-language student. CORD (Jing et al., 2026) aligns audio-conditioned reasoning with its text-conditioned counterpart within a unified model, leveraging the text modality as an internal teacher and applying importance-aware weighting to prioritize early and semantically critical tokens.

However, these methods primarily target settings where the teacher's additional information is expected to remain recoverable from the student's modality. Our setting considers a different regime, where part of the teacher's advantage may arise from information that is not fully recoverable from the student's observation. Consider a depth-privileged teacher and an RGB-only student. Some teacher-generated tokens are directly grounded in depth information that the student cannot infer

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

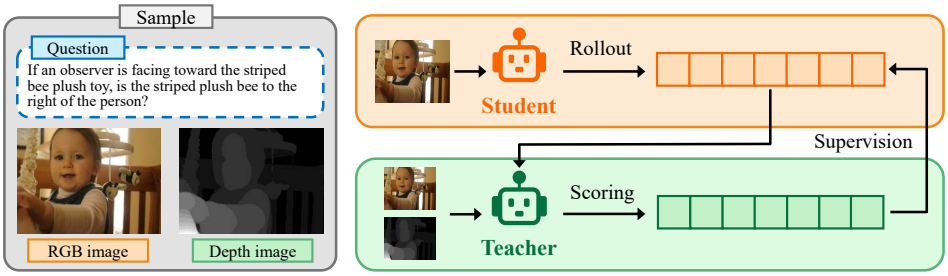

Figure 1: An example of our cross-modal OPD scenario.

from appearance. Others arise when depth helps the teacher interpret or reason over visual content that is already present in the RGB image. We call these *privilege-grounded tokens* and *privilege-mediated tokens*, respectively. The former carries information unavailable to the student, while the latter captures reasoning that remains grounded in student-visible content but is facilitated by the privilege modality. Existing methods upweight both identically, diluting the transferable signal while injecting supervision the student cannot reproduce.

To address this challenge, we propose LAMP, a cross-modal on-policy distillation framework that weighs tokens with learnability attribution and lookahead divergence. Our key insight is that the teacher's per-token advantage in this scenario requires two complementary modulations. The first is **learnability-gated modality attribution**: it builds a self-calibrating gate with controlled counterfactual modality perturbations to distinguish privilege-grounded from privilege-mediated tokens. This allows the distillation process to suppress supervision that the student's modality cannot support while preserving transferable reasoning. The second is **horizon-conditioned divergence weighting**: it dynamically weighs each token by its downstream influence, so the distillation concentrates learning attention on positions that shape subsequent reasoning.

We evaluate LAMP on two scenarios with six benchmarks, regarding depth image and video as the privilege modality respectively. These benchmarks cover various settings from embodied reasoning, egocentric spatial relations, to visual understanding and visual logical reasoning. Across these benchmarks, LAMP achieves relative improvements of up to 6.1% over the best-performing baseline in the depth-to-RGB setting and 10.7% in the video-to-RGB setting. Our contributions can be summarized as follows:

- We identify and formalize the problem where the teacher's per-token advantage is grounded in a privilege modality whose information cannot be fully recovered from the student's modality, distinguishing sensor tokens from reasoning tokens.
- We propose a self-calibrating gate built from controlled counterfactual modality perturbations that separates sensor tokens from reasoning tokens, suppressing supervision the student's modality cannot support.
- We adopt a weighting scheme that dynamically adjusts each token's learning weight by its downstream influence, concentrating learning attention on positions that shape subsequent reasoning.
- Extensive evaluations on scenarios regarding depth image or video as the privilege modality demonstrate that LAMP achieves up to 10.7% performance gain against the strongest baseline.

2 Background and Motivation

2.1 Standard Cross-Modal OPD and Its Blind Spot

On-policy distillation (OPD) trains a student policy π_θ on its own rollouts while a teacher policy π_T scores the tokens the student actually samples. Following the GKD formulation (Agarwal et al., 2024), for an input x and a student-generated sequence y, the per-token reverse KL is

$$D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid x, y_{<t}) \,\|\, \pi_T(\cdot \mid x, y_{<t})\big), \tag{1}$$

where $y_{<t}$ is the prefix before token y_t. The sequence-level objective aggregates this signal over student trajectories, with every token receiving the same weight:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\Big[\sum_{t=1}^{|y|} D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid x, y_{<t}) \,\|\, \pi_T(\cdot \mid x, y_{<t})\big)\Big]. \tag{2}$$

Cross-modal OPD extends this recipe to settings where the teacher accesses a modality M_T that the student does not possess. Existing methods (Fu et al., 2026; Cao et al., 2026) primarily target settings where the teacher's advantage is *reasoning-level*, i.e., the underlying information is expected to remain recoverable from the student's modality. However, in privilege-modality settings, the teacher may benefit from information that is not recoverable from the student modality (Chen et al., 2024). As illustrated in Figure 2, token-level supervision can arise from either recoverable information (Type A) or privilege information unavailable to the student (Type B). In the former case, the teacher provides the same top-1 token choice as the student. But in the latter case, the teacher with privilege input tries to directly jump to a conclusion, while model with only RGB image input tries to start reasoning about the spatial relationship. These two cases are indistinguishable in the standard token-level OPD objective, although only the former provides supervision that the student can potentially reproduce.

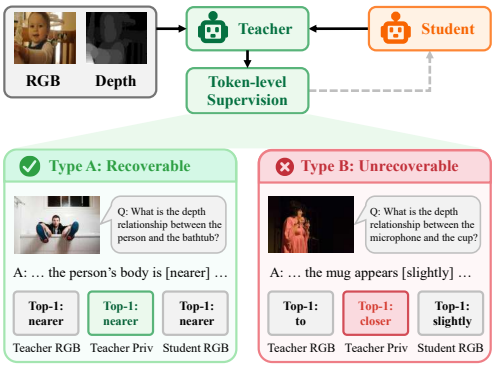

Figure 2: Examples of recoverable and unrecoverable token supervision in cross-modal OPD.

Therefore, we identify two types of teacher-generated tokens according to the source of their predictive advantage. *Privilege-grounded tokens* are tokens whose prediction is primarily supported by information contained in the privilege modality and unavailable from the student modality alone. *Privilege-mediated tokens* are tokens whose prediction is grounded in student-visible content, while the privilege modality helps the teacher interpret that content or carry out the reasoning process.

2.2 Counterfactual Learnability Attribution

Given the distinction between privilege-grounded and privilege-mediated tokens, the next question is how to identify the source of a token's predictive advantage. This source is not directly observable from the teacher's internal computation: the teacher does not reveal which modality supports its prediction at a particular token, while the student does not have access to the privilege modality. Although the source of a token's predictive advantage is not directly observable, its dependence on each modality can be probed through controlled input interventions. If a token prediction relies on information from a particular modality, altering that modality while keeping the other factors fixed should change the teacher's next-token distribution. We therefore use the teacher's sensitivity to such interventions as an observable proxy for modality dependence.

Specifically, we consider two complementary interventions. The first removes the privilege modality while preserving the student modality, probing how much the teacher's prediction depends on privilege information. The second perturbs the student modality while preserving the privilege modality, probing how much the prediction remains sensitive to information available to the student. These two interventions are considered jointly because sensitivity to one modality alone does not determine the source of the teacher's advantage. For example, a change in the student modality may alter the teacher's prediction simply because the teacher relies on visual context when processing the privilege modality. Comparing the teacher's responses to both interventions therefore provides a more informative signal for distinguishing privilege-grounded supervision from privilege-mediated supervision.

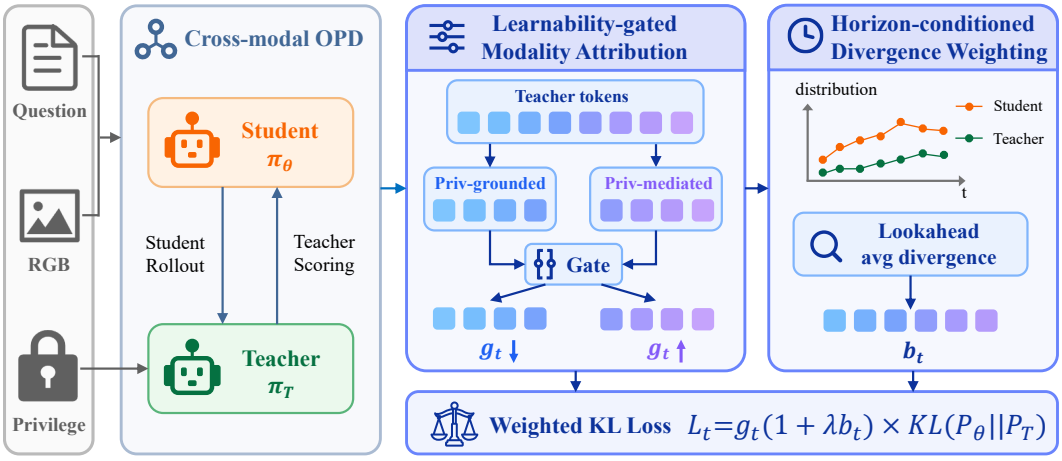

Figure 3: An overview of LAMP. We first roll out the student policy with the student modality and text prompt, followed by teacher scoring with privilege input. Learnability-gated modality attribution and horizon-conditioned divergence weighting are then used to construct the per-token weight for the OPD objective.

2.3 Horizon-Conditioned Divergence Weighting

The learnability attribution determines how much supervision should be retained at each token, but it does not determine where the distillation signal should be concentrated. In reasoning tasks, tokens are not equally consequential: some positions are followed by substantial changes in the subsequent reasoning trajectory, while others are followed by relatively stable behavior. Related work on chain-of-thought distillation has likewise shown that rationale tokens vary in importance and can benefit from token-specific weighting (Feng et al., 2024; Kim & Baek, 2026; Le et al., 2025). Treating all transferable tokens with equal weight can therefore allocate learning capacity to positions that have limited downstream relevance.

This raises a second, complementary question: where should the distillation signal be concentrated? A token's local teacher-student disagreement does not necessarily reflect its downstream relevance, since a token can have low local divergence while being followed by increasing divergence over subsequent reasoning steps, suggesting greater downstream relevance than its local signal suggests. Therefore, the local signal only provides limited information about how much learning attention should be allocated to that position. Taking this into account, we propose to look beyond the current token. By assessing how teacher-student disagreement evolves over the subsequent reasoning trajectory, we can obtain additional information about a token's relevance to the ongoing reasoning process.

3 LAMP Design

We propose LAMP, a cross-modal on-policy distillation framework for privilege-modality-to-image transfer in MLLMs. As illustrated in Figure 3, we first perform learnability-gated modality attribution to distinguish transferable supervision from supervision grounded in the privilege modality. We then estimate the downstream relevance of each token from future teacher–student divergence. The two signals are combined into a per-token weight, which modulates the reverse-KL objective of on-policy distillation (OPD).

3.1 Problem Formulation

We consider cross-modal on-policy distillation where a student policy π_θ operates on a student modality M_S, while a teacher policy π_T additionally accesses a privilege modality M_T. Let $y = (y_1, \ldots, y_{|y|})$ denote a trajectory sampled from the student $\pi_\theta(\cdot \mid M_S)$, and let $y_{<t}$ denote the prefix before token y_t. The student and teacher distributions are

$$\pi_\theta(\cdot \mid M_S, y_{<t}) \qquad \text{and} \qquad \pi_T(\cdot \mid M_S, y_{<t}, M_T). \tag{3}$$

Standard OPD applies the same weight to every token. We instead introduce a token-dependent weight w_t to modulate the distillation signal. The goal is to suppress supervision that depends on the privilege modality while emphasizing transferable supervision that is consequential to subsequent reasoning.

3.2 Learnability-Gated Modality Attribution

We determine whether the teacher's predictive advantage stems from the privilege modality itself or from information that is also available to the student through controlled modality interventions. To disentangle these two possibilities, we construct three teacher configurations with shared frozen model weights: a **full teacher** P_{full} receiving both M_S and M_T, an **ablated teacher** P_{abl} receiving only M_S, and a **perturbed teacher** P_{pert} receiving a perturbed student modality $\tilde{M}_S$ together with M_T. At each token t, we compute

$$\Delta_{\text{priv}}(t) = D_{\text{KL}}\left(P_{\text{full}}(\cdot \mid M_S, M_T, y_{<t}) \,\|\, P_{\text{abl}}(\cdot \mid M_S, y_{<t})\right), \tag{4}$$

$$\Delta_{\text{stud}}(t) = D_{\text{KL}}\left(P_{\text{full}}(\cdot \mid M_S, M_T, y_{<t}) \,\middle\|\, P_{\text{pert}}(\cdot \mid \tilde{M}_S, M_T, y_{<t})\right). \tag{5}$$

Δ_{priv} measures sensitivity to the privilege modality, while Δ_{stud} measures sensitivity to information available through the student modality. We normalize both signals using batch-level exponential moving averages:

$$\hat{\Delta}_{\text{priv}}(t) = \frac{\Delta_{\text{priv}}(t)}{\Delta_{\text{priv}}(t) + \mu_{\text{priv}}}, \qquad \hat{\Delta}_{\text{stud}}(t) = \frac{\Delta_{\text{stud}}(t)}{\Delta_{\text{stud}}(t) + \mu_{\text{stud}}}, \tag{6}$$

where μ_{priv} and μ_{stud} are EMA-smoothed batch means. We then define the attribution gate as

$$g_t = 1 + \hat{\Delta}_{\text{priv}}(t)\left(2\hat{\Delta}_{\text{stud}}(t) - 1\right). \tag{7}$$

The gate takes values in $[0, 2]$. Tokens strongly dependent on the privilege modality but insensitive to the student modality obtain low g_t, whereas tokens sensitive to information available in both modalities receive larger values.

3.3 Horizon-Conditioned Divergence Weighting

The attribution gate determines the learnability of teacher supervision, but does not distinguish tokens according to their downstream relevance. We therefore estimate the relevance of token t from the divergence that unfolds over the following k steps. Let $k_t = \min(k, |y| - t)$ and define

$$A_t = \frac{1}{k_t} \sum_{i=1}^{k_t} \left|\log \pi_\theta(y_{t+i} \mid s_{t+i}) - \log \pi_T(y_{t+i} \mid s_{t+i})\right|. \tag{8}$$

Large A_t indicates that the token is followed by sustained teacher–student divergence and is therefore potentially consequential to the subsequent trajectory. We normalize A_t within each sequence to obtain

$$b_t = \text{clamp}\left(\frac{A_t - \min(A)}{\max(A) - \min(A) + \epsilon}, 0, 1\right). \tag{9}$$

Rather than adding b_t directly to the attribution gate, we use it as a multiplicative modulation. This preserves the role of g_t as the base learnability gate: downstream relevance can amplify learnable supervision, but cannot restore supervision that is suppressed by the attribution gate.

3.4 Combined Weight and OPD Objective

We combine the two signals as

$$w_t = g_t \left(1 + \lambda b_t\right), \tag{10}$$

where $\lambda \geq 0$ controls the strength of horizon-conditioned reweighting. When $\lambda = 0$, the method reduces to attribution-only weighting, while $g_t \approx 0$ suppresses supervision dominated by privilege information regardless of its downstream relevance.

Following OPD, we optimize the reverse KL divergence from the student policy to the teacher policy with the proposed token-dependent weight:

$$\mathcal{L}(\theta) = \mathbb{E}_{y \sim \pi_\theta(\cdot|M_S)} \left[\sum_{t=1}^{|y|} w_t D_{\mathrm{KL}} \left(\pi_\theta(\cdot \mid M_S, y_{<t}) \,\|\, \pi_T(\cdot \mid M_S, y_{<t}, M_T) \right) \right]. \tag{11}$$

The teacher distributions and weighting signals are detached during optimization, so gradients flow only through the student policy.

4 Experiments

4.1 Experimental setup

Datasets and models. We evaluate LAMP in two cross-modal post-training scenarios, where the privilege modality is available to the teacher during training but is not accessible to the student at inference time. In the depth-to-RGB setting, the student receives RGB images while the teacher additionally accesses depth images. We use the In-the-Wild-o365 split released with HiSpatial (Liang et al., 2026) for training and evaluate the resulting models on ERQA (Gemini Robotics Team et al., 2025), EmbSpatial-Bench (Du et al., 2024), and RoboSpatial (Song et al., 2025). In the video-to-RGB setting, the student operates on RGB inputs while the teacher additionally accesses video information. We use FineVideo (Farré et al., 2024) for training and evaluate on LogicVista (Xiao et al., 2024), VisuLogic (Xu et al., 2026), and V*Bench (Wu & Xie, 2024) (without tool usage). We adopt Qwen3-VL-8B-Thinking as the frozen teacher and Qwen3-VL-2B-Thinking and Qwen3-VL-4B-Thinking as student models (Bai et al., 2025).

Implementation. We implement LAMP using verl (Sheng et al., 2025) with $4\times$NVIDIA RTX PRO 6000 Blackwell GPUs. For all scenarios and student models, we use a batch size of 32, a group size of 4, a learning rate of $1\mathrm{e}{-6}$, and 125 training steps. For GRPO, we use a group size of 8. The maximum rollout reasoning length is set to 8192 tokens for the depth-to-RGB scenario and 2048 tokens for the video-to-RGB scenario.

Baselines and evaluations. We compare LAMP with the corresponding unmodified Qwen3-VL checkpoint (Base Model) (Bai et al., 2025), GRPO (Shao et al., 2024), OPD (Agarwal et al., 2024), X^3-OPD (Fu et al., 2026), and CORD (Jing et al., 2026). All post-training methods use the same training data within each scenario. We follow the official Qwen3-VL evaluation settings (Bai et al., 2025) for all models and benchmarks and report the final accuracy in percentage. For ERQA, RoboSpatial, VisuLogic and V*Bench, we report avg@4 accuracy, and for EmbSpatial-Bench and LogicVista, we report avg@1 accuracy.

4.2 Main results

Table 1 presents the main results for privilege-modality-to-RGB transfer. The experiment evaluates whether an RGB-only student can benefit from a teacher that has access to additional depth or video information, and compares LAMP with both standard post-training and existing cross-modal on-policy distillation methods.

Overall performance. LAMP achieves the highest average performance for both Qwen3-VL-2B-Thinking and Qwen3-VL-4B-Thinking. For the 2B student, the average score improves from

Table 1: Performance comparison across scenarios, base models and benchmarks. Best results are emphasized in **bold** and second best in underlined.

Method	Depth→RGB			Video→RGB			Average
	ERQA	EmbSpatial	RoboSpatial	LogicVista	VisuLogic	V*	
Qwen3-VL-2B-Thinking							
Base Model	39.8	71.5	41.2	40.7	25.3	69.6	48.0
GRPO	40.0 (+0.2)	71.5 (0.0)	44.0 (+2.8)	45.1 (+4.4)	27.6 (+2.3)	70.8 (+1.2)	49.8 (+1.8)
OPD	40.3 (+0.5)	72.5 (+1.0)	41.2 (0.0)	44.0 (+3.3)	23.7 (-1.6)	69.1 (-0.5)	48.5 (+0.5)
X^3-OPD	39.8 (0.0)	72.2 (+0.7)	43.3 (+2.1)	42.0 (+1.3)	24.7 (-0.6)	70.4 (+0.8)	48.7 (+0.7)
CORD	40.0 (+0.2)	72.8 (+1.3)	42.4 (+1.2)	46.0 (+5.3)	27.5 (+2.2)	70.6 (+1.0)	49.9 (+1.9)
LAMP	**41.7 (+1.9)**	**72.9 (+1.4)**	**46.7 (+5.5)**	**47.2 (+6.5)**	**28.3 (+3.0)**	**71.4 (+1.8)**	**51.4 (+3.4)**
Qwen3-VL-4B-Thinking							
Base Model	40.7	77.3	51.9	55.8	27.8	78.7	55.4
GRPO	41.9 (+1.2)	77.1 (-0.2)	55.5 (+3.6)	54.5 (-1.3)	26.8 (-1.0)	79.3 (+0.6)	55.9 (+0.5)
OPD	42.5 (+1.8)	77.5 (+0.2)	56.0 (+4.1)	55.8 (0.0)	28.0 (+0.2)	79.3 (+0.6)	56.5 (+1.2)
X^3-OPD	42.0 (+1.3)	77.5 (+0.2)	55.5 (+3.6)	56.2 (+0.4)	25.9 (-1.9)	79.8 (+1.1)	56.2 (+0.8)
CORD	42.0 (+1.3)	76.0 (-1.3)	56.2 (+4.3)	57.0 (+1.2)	28.0 (+0.2)	79.2 (+0.5)	56.4 (+1.0)
LAMP	**44.0 (+3.3)**	**77.8 (+0.5)**	**58.5 (+6.6)**	**58.1 (+2.3)**	**31.0 (+3.2)**	**80.2 (+1.5)**	**58.3 (+2.9)**

48.0 for the base model to 51.4 with LAMP, while for the 4B student it improves from 55.4 to 58.3. The consistent improvements across the two student scales indicate that LAMP benefits from privilege-modality supervision without relying on a particular student capacity. Moreover, the gains are observed in both transfer scenarios, suggesting that the proposed weighting mechanism is applicable to qualitatively different forms of privilege information, rather than being specialized to either geometric or temporal cues.

Comparison with existing distillation methods. Standard OPD and existing cross-modal distillation methods can improve over the base model, but their gains are less consistent across benchmarks. LAMP achieves the best result on all six benchmarks for both student models, and the strongest relative improvements reach 6.1% in the depth-to-RGB setting and 10.7% in the video-to-RGB setting. These results suggest that simply exposing the student to a stronger teacher is not sufficient when part of the teacher's advantage comes from information unavailable to the student. Instead, selectively weighting the teacher signal allows LAMP to retain useful supervision while avoiding uniform imitation of predictions that depend directly on the privilege modality.

Beyond the overall gains, LAMP exhibits a consistently positive improvement pattern across individual benchmarks. Existing methods often show uneven performance, while LAMP improves over the corresponding base model on all six benchmarks for both student models, maintaining gains over the strongest competing methods. This consistency is notable given the diversity of the evaluated benchmarks, which cover different forms of visual, spatial, and logical reasoning. These results suggest that selectively filtering and reweighting privilege-modality supervision enables more broadly transferable improvements for RGB-only reasoning.

4.3 Mechanistic Analysis of LAMP

To better understand how LAMP reshapes the distillation signal, we analyze how its token-level weights are distributed and whether the learnability gate g_t and horizon-conditioned bonus b_t provide complementary signals.

As shown in Fig. 4(a), standard OPD assigns a uniform weight of 1 to every token, whereas LAMP produces a more diverse distribution. A large portion of tokens receive weights below 1, while a smaller subset is assigned larger weights, indicating that LAMP selectively reallocates supervision across the reasoning trajectory. This reallocation suggests that the distillation signal is concentrated on tokens that are estimated to provide more useful learning signals, while the influence of less informative tokens is correspondingly reduced.

To further examine the roles of the two factors, we partition rollout tokens into four groups according to whether g_t and b_t are above or below their respective medians, and train variants using only the tokens in each group under the same training budget. As shown in Fig. 4(b), the group with both g_t and b_t above their medians achieves the best average performance (52.7), while the group with both below their medians performs the worst (51.1). The two mixed groups obtain intermediate

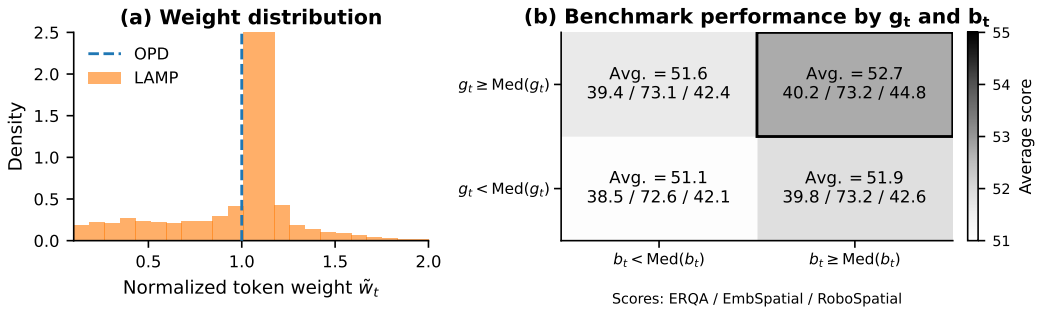

Figure 4: Mechanistic analysis of LAMP's token-level weighting. (a) Distribution of normalized token weights for standard OPD and LAMP. (b) Benchmark performance obtained by partitioning tokens according to g_t and b_t.

performance (51.9 and 51.6). These results suggest that learnability and downstream relevance provide complementary signals for identifying useful supervision.

Finally, Fig. 5 provides a token-level example of this selective weighting. Within the same reasoning span, relational tokens such as 'behind' and 'maybe' receive relatively large weights of 1.79 and 1.73, whereas 'next' receives a much smaller weight of 0.39. In this example, tokens associated with the evolving relational reasoning receive stronger supervision, while less consequential tokens are assigned lower weights. This illustrates that LAMP assigns different supervision strengths to tokens within the same reasoning trajectory.

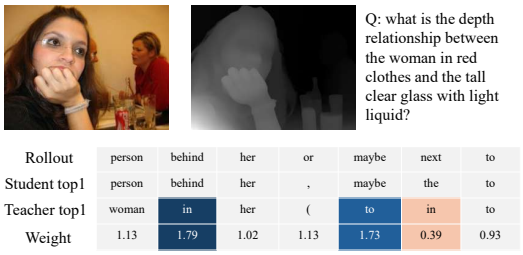

Figure 5: Token-level case study of LAMP on a representative rollout.

4.4 Self-Distillation Results

We further evaluate LAMP in a self-distillation setting, where Qwen3-VL-4B-Thinking serves as both the teacher and student. We include OPSD (Zhao et al., 2026) as a baseline for on-policy self-distillation.

As shown in Table 2, LAMP consistently improves over both the base model and OPSD across all six benchmarks. On the depth benchmarks, LAMP improves the average score from 56.6 for the base model and 58.0 for OPSD to 59.5. On the video benchmarks, LAMP achieves 56.0, compared with 54.1 for the base model and 54.4 for OPSD. The consistent gains over OPSD suggest that the improvement is not solely due to self-distillation, but also comes from selectively reweighting the supervision.

Table 2: Self-distillation results with Qwen3-VL-4B-Thinking as the model.

Method	ERQA	EmbSpatial	RoboSpatial
Base Model	40.7	77.3	51.9
OPSD	42.0 (+1.3)	77.5 (+0.2)	54.5 (+2.6)
LAMP	**42.8** (+2.1)	**77.8** (+0.5)	**57.9** (+6.0)
Method	LogicVista	VisuLogic	V*
Base Model	55.8	27.8	78.7
OPSD	56.5 (+1.0)	27.6 (-0.2)	79.0 (+0.3)
LAMP	**57.6** (+1.8)	**30.9** (+1.1)	**79.5** (+0.8)

4.5 Ablation Study

We ablate the two main components of LAMP in the depth-to-RGB setting using Qwen3-VL-2B-Thinking. Removing either learnability-gated modality attribution or horizon-conditioned divergence weighting consistently reduces performance across the three benchmarks. The degradation is larger without attribution, especially on RoboSpatial, in-

Table 3: Ablation study on depth→RGB distillation using Qwen3-VL-2B-Thinking as student.

Method	ERQA	EmbSpatial	RoboSpatial
LAMP w/o attribution	40.5	72.5	43.0
LAMP w/o weighting	41.0	72.3	45.9
LAMP	**41.7**	**72.9**	**46.7**

Table 4: Capability preserving evaluation with Qwen3-VL-2B-Thinking as the student model.

Method	Depth→RGB			Video→RGB		
	LogicVista	VisuLogic	V*	ERQA	EmbSpatial	RoboSpatial
Base Model	40.7	25.3	69.6	39.8	71.5	41.2
GRPO	40.5 (-0.2)	25.4 (+0.1)	70.2 (+0.6)	39.9 (+0.1)	71.2 (-0.3)	41.6 (+0.4)
OPD	41.0 (+0.3)	25.2 (-0.1)	70.2 (+0.6)	40.4 (+0.6)	71.6 (+0.1)	39.8 (-1.4)
X^3-OPD	41.0 (+0.3)	24.3 (-1.0)	68.6 (-1.0)	39.8 (0.0)	71.1 (-0.4)	40.3 (-0.9)
CORD	40.3 (-0.4)	26.4 (+1.1)	70.0 (+0.4)	39.5 (-0.3)	70.7 (-0.8)	40.8 (-0.4)
LAMP	40.9 (+0.2)	26.5 (+1.2)	70.4 (+0.8)	40.2 (+0.4)	72.0 (+0.5)	41.4 (+0.2)

dicating that filtering supervision according to its learnability is particularly important. Removing weighting also causes a consistent drop, supporting its complementary role in emphasizing more consequential reasoning positions.

4.6 Capability preserving results

The main experiments study privilege-modality transfer, while we additionally examine whether post-training affects the model's existing capabilities. Table 4 compares Qwen3-VL-2B-Thinking before and after training with different post-training methods. GRPO, OPD, X^3-OPD, and CORD show mixed changes across the six benchmarks, with gains on some tasks accompanied by degradations on others. In comparison, LAMP remains close to the base model overall, while improving EmbSpatial, RoboSpatial, and VisuLogic and maintaining comparable performance on the remaining benchmarks. These results suggest that LAMP can enhance the targeted capabilities while largely preserving the model's existing capabilities.

5 Related Works

LLM post-training and on-policy distillation. LLM post-training includes supervised fine-tuning, preference optimization, reinforcement learning, and distillation. On-policy distillation (OPD) has recently emerged as a promising paradigm, training students on their rollouts with dense token-level supervision from teachers, typically via reverse KL (Agarwal et al., 2024). Recent work improves this objective through entropy-based balancing (Jin et al., 2026; Lee et al., 2026), adaptive per-token KL direction (Jung et al., 2025), tokenizer-mismatch handling (Niu et al., 2026), and black-box distillation without teacher logits (Ye et al., 2025).

Multimodal and cross-modal on-policy distillation. Multimodal OPD extends this paradigm to vision-language settings. H-OPD (Yin et al., 2026) arbitrates supervision from heterogeneous teachers, OnPoKD (Zhang et al., 2026) constructs adaptive targets from reliability and disagreement, and ViCuR (Tian et al., 2026) treats visual cues as recoverable privilege and jointly trains a cue recovery module. Related work also explores on-policy self-distillation with a privileged teacher (Li et al., 2026; Wang et al., 2026c;b;a; Luo et al., 2026). Cross-modal OPD transfers supervision across modalities, including X^3-OPD (Fu et al., 2026), X-OPD (Cao et al., 2026), and CORD (Jing et al., 2026). These methods generally assume that the teacher's advantage is reasoning-level and recoverable from the student's modality, whereas our work targets information-level advantages that require per-token attribution before imitation.

6 Conclusion

We propose LAMP, a learnability- and horizon-aware framework for cross-modal on-policy distillation from privilege modalities to RGB-only MLLMs. LAMP addresses the limitation that teacher supervision may arise either from student-visible content or privilege information unavailable to the student, by introducing counterfactual learnability attribution and horizon-conditioned divergence weighting to respectively assess supervision learnability and downstream relevance. Experiments on depth-to-RGB and video-to-RGB transfer show that LAMP consistently improves over the strongest baselines, and remains effective under self-distillation. These results demonstrate the effectiveness of selectively weighting privilege-modality supervision according to both its learnability and its role in subsequent reasoning.

AI DISCLOSURE

In this work, we used generative AI tools to aid or polish writing, retrieve or find related works, and implement code for our method. We have not used generative AI tools to generate research ideas or synthetic dataset, prove mathematical theorems, or other tasks with required disclosure. We have reviewed all AI-assisted work. We have checked LLM-searched related works and verified LLM-generated code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Di Cao, Dongjie Fu, Hai Yu, Siqi Zheng, Xu Tan, and Tao Jin. X-OPD: Cross-modal on-policy distillation for capability alignment in speech LLMs. *arXiv preprint arXiv:2603.24596*, 2026.

Mengxi Chen, Fei Zhang, Zihua Zhao, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Probabilistic conformal distillation for enhancing missing modality robustness. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1142.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, et al. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 346–355, 2024.

Miquel Farré, Andres Marafioti, Lewis Tunstall, Leandro von Werra, Pedro Cuenca, and Thomas Wolf. FineVideo. Hugging Face FineVideo dataset, 2024.

Kaituo Feng, Changsheng Li, Xiaolu Zhang, Jun Zhou, Ye Yuan, and Guoren Wang. Keypoint-based progressive chain-of-thought distillation for LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 13241–13255, 2024.

Dongjie Fu, Di Cao, Xize Cheng, Zihan Zhang, Wenxu Jia, Yifu Chen, Shengpeng Ji, Yu Zhang, and Tao Jin. X^3-OPD: Distilling reasoning into large audio-language models via on-policy alignment. *arXiv preprint arXiv:2607.21550*, 2026.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, et al. Gemini robotics: Bringing AI into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

GLM-5 Team, Aohan Zeng, Xin Lv, Zhenyu Hou, et al. GLM-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Hu Jing, Danxiang Zhu, Xianlong Luo, Dan Zhang, Shuwei He, Yishu Lei, Shikun Feng, Hai-Tao Zheng, Jingzhou He, Yu Sun, et al. CORD: Bridging the audio–text reasoning gap via weighted on-policy cross-modal distillation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 31602–31612, 2026.

Seongryong Jung, Suwan Yoon, DongGeon Kim, and Hwanhee Lee. ToDi: Token-wise distillation via fine-grained divergence control. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8091, 2025. doi: 10.18653/v1/2025.emnlp-main.409.

Sai Srinivas Kancheti, Aditya Sanjiv Kanade, Vineeth N Balasubramanian, and Tanuja Ganu. Chain-of-thought degrades visual spatial reasoning capabilities of multimodal LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 862–876, 2026.

Minsang Kim and Seung Jun Baek. Explain in your own words: Improving reasoning via token-selective dual knowledge distillation. In *International Conference on Learning Representations*, 2026.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, et al. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Tue Le, Hoang Tran, Quyen Tran, Linh Ngo, Mehrtash Harandi, and Trung Le. Token-level self-play with importance-aware guidance for large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4030.

Chia-Hsuan Lee, Zelei Cheng, Yu Wang, Renkun Ni, Sambit Sahu, Shi-Xiong Zhang, and William Campbell. SEAD: Competence-aware on-policy distillation via entropy-guided supervision. *arXiv preprint arXiv:2606.28562*, 2026.

Pengyu Li, Zhitao Gao, Lingling Zhang, Muye Huang, Yuanming Li, Fangzhi Xu, and Jun Liu. Visual-OPSD: Cross-modal on-policy self-distillation for efficient unified multimodal reasoning. *arXiv preprint arXiv:2606.18974*, 2026.

Huizhi Liang, Yichao Shen, Yu Deng, Sicheng Xu, Zhiyuan Feng, Tong Zhang, Yaobo Liang, and Jiaolong Yang. HiSpatial: Taming hierarchical 3D spatial understanding in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Haocheng Luo, Jiahui Liu, Ruicheng Zhang, Zhizhou Zhong, Jiaqi Huang, Zunnan Xu, Quan Shi, Jun Zhou, Shuiyang Mao, Wei Liu, et al. Learning visual spatial planning from symbolic state via modality-gap-aware self-distillation. *arXiv preprint arXiv:2606.06076*, 2026.

Yifan Niu, Han Xiao, Dongyi Liu, Zelong Wang, Dihong Gong, Yasheng Wang, and Jia Li. Breaking the tokenizer barrier: On-policy distillation across model families. *arXiv preprint arXiv:2606.09456*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient RLHF framework. In *Proceedings of the 20th European Conference on Computer Systems*, pp. 1279–1297, 2025. doi: 10.1145/3689031.3696075.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. RoboSpatial: Teaching spatial understanding to 2D and 3D vision-language models for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15768–15780, 2025.

Kanghui Tian, Siyuan Liu, Ziang Yan, Sheng Xia, Shuai Dong, and Yi Wang. ViCuR: Visual cues as recoverable privilege for multimodal on-policy distillation. *arXiv preprint arXiv:2606.05718*, 2026.

Kaishen Wang, Dongdi Zhao, Yijun Liang, Dingqiang Ye, Ruibo Chen, Heng Huang, and Di Fu. Where to look matters: On-policy self-distillation for long-video understanding. *arXiv preprint arXiv:2608.25356*, 2026a.

Shuai Wang, Daoan Zhang, Zhe Tang, Hao Cheng, and Jiaheng Wei. Self-boosting vision-language models with noisy student on-policy self-distillation. *arXiv preprint arXiv:2607.23125*, 2026b.

Sihan Wang, Xiyao Liu, Lianqing Liu, and Zhi Han. Seeing before reasoning: Decoupling perception and reasoning for shortcut-resilient multimodal on-policy self-distillation. *arXiv preprint arXiv:2606.19120*, 2026c.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. LogicVista: Multimodal LLM logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. VisuLogic: A benchmark for evaluating visual reasoning in multi-modal large language models. In *International Conference on Learning Representations*, 2026.

Wenda Xu, Rujun Han, Zifeng Wang, Long T. Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations*, 2025.

Lecheng Yan, Ruizhe Li, Jiahui Geng, Qing Li, Minghao Wu, Zhanyu Wang, Wenxi Li, Tianbo Ji, Shaochen Jiang, and Chenyang Lyu. On the taxonomy, tasks, and open-challenges for multimodal large language models. In *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2518–2525. IEEE, 2025.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models. *arXiv preprint arXiv:2511.10643*, 2025.

Qixiang Yin, Huanjin Yao, Yuchen Cai, Jianghao Chen, Ziyi Wang, Min Yang, Fei Su, and Zhicheng Zhao. H-OPD: Confidence aware heterogeneous multi-teacher multimodal on-policy distillation. *arXiv preprint arXiv:2607.02592*, 2026.

Hongyuan Zhang, Xianda Guo, Yanlun Peng, Qianlong Yang, Yubin Guo, Pinhan Fu, Mulin Chen, Xiaozhen Qiao, and Ping Luo. On-policy distillation for vision-language model adaptation, an effective paradigm on low-quality multimodal data. *arXiv preprint arXiv:2609.10321*, 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

A Appendix Outline

In this appendix, we report the following sections as supplementary materials for the main paper.

- Appendix B provides detailed descriptions of training datasets, benchmarks, and baselines.
- Appendix C introduces the details for training.
- Appendix D provides hyperparameter sensitivity test results.
- Appendix E demonstrates an extra case to understand LAMP.

B Datasets, Benchmarks, and Baselines

This appendix provides additional details about the training data, evaluation benchmarks, and comparison methods used in Section 4. The two experimental settings share the same principle: the teacher can access an additional modality during post-training, whereas the student receives RGB input only and retains the same RGB-only interface at inference time.

B.1 Training Datasets

In-the-Wild-o365. In-the-Wild-o365 is a large-scale training split released with HiSpatial Liang et al. (2026). It is built from real-world images in Objects365 and augments them with metric geometric information and multi-level spatial question–answer annotations. The annotations cover spatial relations at different levels, including relative position, distance and depth comparison, object size, orientation, and perspective taking.

FineVideo. FineVideo is an open video-understanding dataset containing more than 43,000 videos and approximately 3,400 hours of content Farré et al. (2024). In addition to the videos, it provides temporally localized metadata such as scene descriptions, narrative information, transcripts, and question–answer annotations. This combination makes it suitable for constructing the video-to-RGB setting: the teacher can integrate evidence distributed across the video, whereas the student is restricted to the RGB input specified by our training protocol. FineVideo also covers a broad range of content domains, reducing the risk that the learned transfer behavior is tied to a narrow visual domain.

B.2 Evaluation Benchmarks

B.2.1 Depth-to-RGB Evaluation

ERQA. Embodied Reasoning Question Answering (ERQA) evaluates multimodal reasoning in real-world embodied and robotic scenarios Gemini Robotics Team et al. (2025). Its questions interleave images and text and are presented in a multiple-choice format. The benchmark covers spatial and world-knowledge reasoning needed for embodied interaction, including reasoning about object relations, actions, trajectories, state, and multiple views.

EmbSpatial-Bench. EmbSpatial-Bench is designed to evaluate the spatial understanding of large vision-language models in embodied environments Du et al. (2024). It is automatically derived from embodied scenes and covers six egocentric spatial relations. Unlike generic image-question answering benchmarks, it emphasizes viewpoint-dependent relations that an embodied agent must infer with respect to its own position.

RoboSpatial. RoboSpatial studies spatial understanding for both 2D and 3D vision-language models in robotics-oriented settings Song et al. (2025). Its evaluation examines spatial relations and affordance-related understanding that are relevant to downstream manipulation. Compared with ERQA and EmbSpatial-Bench, RoboSpatial places greater emphasis on the connection between geometric reasoning and actionable robotic semantics.

B.2.2 VIDEO-TO-RGB EVALUATION

LogicVista. LogicVista evaluates logical reasoning grounded in visual contexts Xiao et al. (2024). It contains 448 human-annotated multiple-choice questions spanning five core task families—deduction, induction, spatial reasoning, numerical reasoning, and mechanical reasoning—which are further divided into nine multimodal capabilities.

VisuLogic. VisuLogic is a vision-centric reasoning benchmark containing 1,000 human-verified problems across six categories, including quantitative shifts, spatial relations, and attribute comparisons Xu et al. (2026). The benchmark is constructed to reduce the effectiveness of language-only shortcuts and therefore requires the model to reason over information expressed in the image itself.

V*Bench. V*Bench evaluates the ability of multimodal models to locate and reason about fine-grained details in high-resolution images Wu & Xie (2024). Its tasks include attribute recognition and spatial-relation reasoning for visually small or difficult-to-localize targets.

B.3 BASELINES

Base Model. The Base Model is the corresponding unmodified Qwen3-VL-Thinking checkpoint Bai et al. (2025). It receives the same RGB-only evaluation input as every trained student but undergoes no task-specific post-training. This comparison quantifies the net effect of each post-training method relative to the original model and helps distinguish transferred capability from capability already present in the pretrained checkpoint.

GRPO. Group Relative Policy Optimization (GRPO) is a reinforcement-learning algorithm that estimates relative advantages from multiple responses sampled for the same prompt, avoiding a separately learned critic model Shao et al. (2024). In our experiments, GRPO serves as a non-distillation post-training baseline. It learns from sequence-level task feedback but does not receive the teacher's dense token-level distribution, allowing us to separate the benefit of privileged-modality distillation from that of reinforcement learning on the same training data.

OPD. Standard on-policy distillation (OPD) trains the student on trajectories sampled from its own policy while the teacher scores the student-visited states Agarwal et al. (2024). We implement OPD with the conventional token-level reverse-KL objective. This provides dense supervision and avoids the trajectory mismatch of distilling only on teacher-generated responses.

X^3-OPD. X^3-OPD is a cross-modal on-policy distillation method originally proposed to transfer reasoning from a text teacher to an audio-language student Fu et al. (2026). The student generates trajectories from its own modality-conditioned policy, and the teacher provides token-level guidance from a matched input in the stronger modality. It is a relevant baseline because it directly tests cross-modal reasoning transfer along student rollouts.

CORD. CORD performs weighted on-policy cross-modal distillation between audio-conditioned and text-conditioned reasoning within a unified model Jing et al. (2026). At the token level, it emphasizes positions with larger cross-modal divergence and earlier positions whose errors may propagate through the subsequent trajectory; at the sequence level, it augments this alignment with a GRPO objective. CORD is therefore a stronger comparison than uniformly weighted OPD because it distinguishes token importance.

Comparison protocol. Within each privileged-modality setting, all post-training methods use the same training data and are evaluated on the same benchmark splits under the official Qwen3-VL evaluation protocol. The teacher may access depth or video during training, but every student is evaluated with RGB input only. Consequently, the comparison focuses on how effectively each objective transfers useful supervision under a fixed inference interface, rather than on giving a baseline additional information at test time.

Table 5: Hyperparameter test results on depth→RGB distillation using Qwen3-VL-2B-Thinking as student. LAMP adopts EMA decay 0.5, $k = 10$, and $\lambda = 0.5$

Method	ERQA	EmbSpatial	RoboSpatial
LAMP EMA decay 0.3	40.7	72.4	41.1
LAMP EMA decay 0.7	41.0	72.3	45.9
LAMP	**41.7**	**72.9**	**46.7**
LAMP $k = 5$	41.5	**73.2**	44.5
LAMP $k = 15$	40.8	72.1	45.8
LAMP	**41.7**	72.9	**46.7**
LAMP $\lambda = 1.0$	40.8	72.8	44.3
LAMP	**41.7**	**72.9**	**46.7**



Q: What is the appearance of the person's hair in the frame?
Option:
A. Long and wavy
B. Curly and light brown
C. Short and dark
D. Bald

Rollout	has	very	short	hair	almost	shaved	or	bald
Student top1	has	a	short	hair	and	no	,	very
Teacher top1	in	short	short	hair	and	shaved	.	very
Weight	1.27	1.53	1.35	1.37	1.30	2.46	1.28	1.16

Figure 6: An extra case study of LAMP on a representative rollout with video as the privilege modality.

C Training Details

For all scenarios and student models, we adopt AdamW optimizer with a learning rate of 1e-6 and warmup step of 0. Sampling hyperparameters are set at temperature$= 1.0$, top-p$= 1.0$ and repetition_penalty$= 1.0$ for teacher and student model. During training, the loss clamping constant is set to 10 and gradient clip is set to 1.0. The hyperparameters are set as $k = 10, \lambda = 0.5$ and EMA decay coefficient as 0.5.

D Hyperparameter sensitivity test

To understand how LAMP performs under different hyperparameter settings, we conduct a hyperparameter sensitivity test here. Table 5 shows that LAMP is reasonably robust to moderate changes in its main hyperparameters. The default EMA decay of 0.5 achieves the best overall performance, while both smaller and larger decay values lead to notable lower scores. Similarly, changing the horizon from $k = 10$ to either 5 or 15 reduces the overall performance, indicating that an intermediate lookahead horizon provides a suitable balance between local and downstream information. Increasing λ from 0.5 to 1.0 also leads to a performance drop, suggesting that overly strong horizon-based reweighting may amplify less reliable divergence signals.

E Extra case study

Fig. 6 provides an extra token-level example of LAMP with video as the privilege modality. Within the same reasoning span, the correctional token ‘shaved’ receives a large weight of 2.46. Although it’s not the top-1 choice of student, this information is recoverable from the single RGB image and therefore the token is assigned with the largest weight in this region.