RGB-OPD: REPRESENTATION-GUIDED BUDGETING FOR ON-POLICY DISTILLATION IN MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) transfers reasoning capabilities from a teacher to a student on student-generated trajectories, providing supervision at the states actually visited during training. However, we find that image-induced changes are distributed much more broadly in teacher representations than in sampled-token output probabilities, leaving many representation-sensitive positions under-emphasized by standard OPD in multimodal reasoning. Motivated by this representation–output mismatch, we propose **RGB-OPD**, a representation-guided budgeting framework for multimodal on-policy distillation. RGB-OPD uses teacher representation sensitivity to rank response positions, keeping distillation entirely in the output space. It applies coarse distributional supervision over a compact support constructed from both teacher and student distributions, and reallocates a fixed supervision budget toward the most representation-sensitive positions. A bounded termination correction further addresses truncated reasoning trajectories. Across five out-of-domain multimodal reasoning benchmarks, RGB-OPD improves Standard OPD by 0.80 percentage points in the Qwen3-VL 8B→2B run and by 0.91 percentage points in the 32B→8B run, with higher point estimates on all five benchmarks at both scales.

1 INTRODUCTION

On-policy distillation (OPD) has emerged as an effective paradigm for transferring reasoning capabilities from large teacher models to smaller students, and has recently been extended from language reasoning to multimodal reasoning (Agarwal et al., 2024; Bousselham et al., 2026; Yuan et al., 2026). Unlike distillation on fixed teacher-generated sequences, OPD trains the student on its own trajectories and queries the teacher at the states it actually visits, reducing the train–inference mismatch of fixed-sequence supervision (Lin et al., 2020; Gu et al., 2024). Recent multimodal extensions further incorporate visual signals to improve reasoning transfer (Liu et al., 2026; Yoon et al., 2026). Despite this progress, token-level supervision in multimodal OPD is still largely determined by what is exposed at the model output.

However, visual dependence inside the teacher is not always reflected with the same strength at the sampled-token output: a response position may undergo substantial image-induced changes in representation space while showing only a weak output-side change. Since standard OPD derives its token-wise learning signal from sampled-token discrepancies, such positions can receive limited supervision, which raises a central question: *how can multimodal OPD identify where additional visual supervision is needed when it is only partially exposed at the output?*

Based on our observation, visual dependence is far more diffuse in teacher representations than at the sampled-token output, and at the 20% selection fraction used by our method, an output-side ranking recovers only 29% of the most representation-sensitive positions. At this selection fraction, teacher representations therefore flag positions that an output-side ranking would otherwise miss, and standard OPD, whose token-wise coefficients track the output side, places only slightly more weight on those positions than expected under an unrelated ranking. Rather than treating teacher representations as direct distillation targets (Yang et al., 2026), we use them to identify where output-space supervision is most needed; Figure 1 contrasts the two uses.

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.

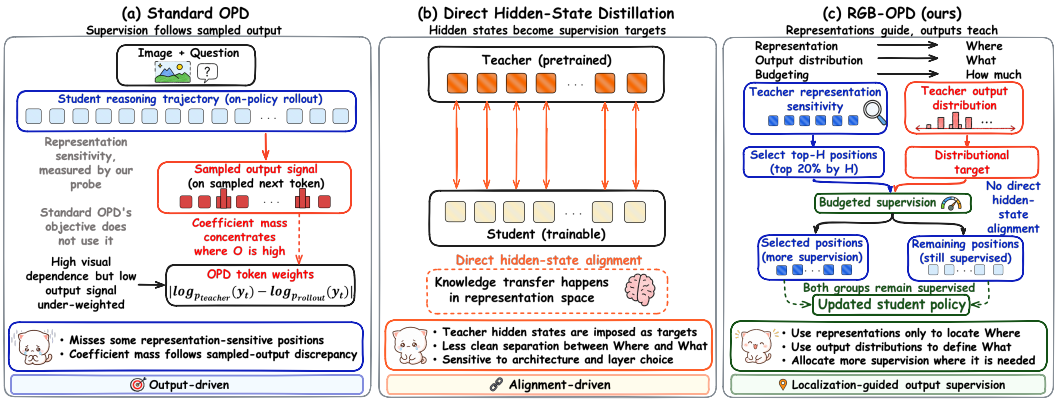


Figure 1: Conceptual comparison of multimodal distillation paradigms. RGB-OPD uses teacher representations for localization rather than direct hidden-state alignment, while distillation remains in the output space.

We therefore propose RGB-OPD, Representation-Guided Budgeting for On-Policy Distillation, which ranks response positions by teacher representation sensitivity and reallocates a fixed distributional supervision budget toward the most sensitive ones. Teacher representations determine where supervision should be concentrated, teacher output distributions determine what the student should match, and budgeting controls how much each group receives. For eligible responses, the remaining positions are down-weighted rather than dropped, so both groups continue to receive distributional supervision. Because a trajectory that reaches the maximum generation length without emitting a termination token is scored as incorrect, the objective also includes a bounded termination correction over the tail of such trajectories. Across five out-of-domain multimodal reasoning benchmarks, RGB-OPD improves Standard OPD by 0.80 percentage points under the Qwen3-VL 8B→2B setting and by 0.91 percentage points under the 32B→8B setting, with higher point estimates on all five benchmarks at both scales.

Our contributions are summarized as follows:

- We identify and quantify a representation–output mismatch in multimodal OPD: output-side signals expose only part of the visual dependence carried by teacher representations.
- We propose RGB-OPD, which uses representation sensitivity to allocate an output-space distillation budget toward visually sensitive positions, together with a bounded termination correction for trajectories that fail to stop.
- We demonstrate improvements over Standard OPD across two teacher–student scales, and compare representation-guided localization with random selection, sampled-token output-guided selection, and leave-one-out component ablations.

2 Related Work

2.1 On-Policy Distillation

On-policy distillation trains a student on trajectories generated by its own policy and queries the teacher at the states actually visited by the student, thereby reducing the train–inference mismatch associated with fixed-sequence distillation (Lin et al., 2020; Agarwal et al., 2024; Song & Zheng, 2026). MiniLLM and GKD develop this paradigm for language-model distillation using student-generated responses and distribution-matching objectives (Gu et al., 2024; Agarwal et al., 2024). More recent work adapts on-policy distillation to multimodal reasoning by incorporating visual information in different forms. VOLD transfers reasoning capabilities from language models to vision-language models (Bousselham et al., 2026), Vision-OPD introduces privileged regional information to supervise a policy operating on full images (Yuan et al., 2026), and Decomposed OPD separates language-prior and visual-grounding components during optimization (Yoon et al., 2026).

VA-OPD further constructs visual advantage from teacher output probabilities and applies it at token and rollout levels (Liu et al., 2026). These studies improve multimodal on-policy distillation primarily through output-side signals, privileged visual information, or optimization-derived quantities. In contrast, we investigate the complementary information exposed by the teacher's internal representations and use representation sensitivity to identify visually dependent positions that are insufficiently characterized by output-side supervision. We discuss recent work on selective, adaptive, and distribution-level distillation in Appendix C.

2.2 Representation Distillation for Multimodal Models

Representation distillation transfers knowledge through intermediate features, hidden states, relational structures, or attention patterns rather than relying solely on output probabilities (Sun et al., 2019; Wang et al., 2020). In multimodal models, representation-based objectives have been used to transfer visual features and cross-modal structure between models: LLaVA-KD distills internal representations and visual-feature relations (Cai et al., 2025), VIRAL aligns multimodal representations with pretrained vision foundation models (Yoon et al., 2025), and HeRA introduces topology-aware alignment over selected attention heads (Caffagni et al., 2026). Representation-level supervision has also been explored under on-policy training; for instance, OPRD aligns teacher and student hidden states along student-generated trajectories (Yang et al., 2026). Rather than treating teacher representations as targets to be reproduced, we use image-induced teacher representation sensitivity to identify response positions where output-space supervision provides an incomplete view of visual dependence, and allocate additional on-policy distillation supervision to these positions.

3 Method

3.1 Preliminary

On-Policy Distillation. On-policy distillation trains a student on its own rollouts and queries the teacher on the prefixes the student actually visits. Given an input x_i, the student generates a response $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,T_i})$. The token-wise distillation signal is

$$A_{i,t} = \log p_T(y_{i,t} \mid x_i, \mathbf{y}_{i,<t}) - \log p_{\text{old}}(y_{i,t} \mid x_i, \mathbf{y}_{i,<t}), \tag{1}$$

where p_T is the teacher policy and p_{old} is the rollout policy. With likelihood ratio

$$\rho_{i,t}(\theta) = \frac{p_\theta(y_{i,t} \mid x_i, \mathbf{y}_{i,<t})}{p_{\text{old}}(y_{i,t} \mid x_i, \mathbf{y}_{i,<t})}, \tag{2}$$

the objective is

$$\mathcal{L}_{\text{OPD}} = -\frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{t=1}^{T_i} m_{i,t}\, \rho_{i,t}(\theta)\, A_{i,t}}{\sum_{t=1}^{T_i} m_{i,t}}, \tag{3}$$

where $m_{i,t}$ masks invalid response tokens. Since $A_{i,t}$ depends only on the fixed teacher and rollout policies, it is constant with respect to the active student parameters. For a fresh on-policy update, $\rho_{i,t}(\theta) = 1$ in value while its gradient remains nonzero, so $A_{i,t}$ acts as the coefficient on the sampled-token log-likelihood gradient.

3.2 Motivation and Analysis

Standard OPD derives its token-wise learning signal from discrepancies observed at the sampled-token output. In multimodal reasoning, however, visual information can substantially affect the teacher's internal computation without being expressed with comparable strength at that output. We investigate this representation–output mismatch from three complementary views: how visual dependence is distributed across response positions, whether output-side sensitivity identifies the same positions as teacher representations, and how standard OPD weights these positions. Figure 2 summarizes the analysis.

For each fixed textual query and student-generated response prefix, we compare the teacher under the original image I_i, a shuffled image I_i^{shuf} that disrupts global visual semantics, and a mildly JPEG-compressed control I_i^{jpg}. Let $\mathbf{h}_{i,t}^{(\ell)}(I)$ denote the teacher representation at layer ℓ and response

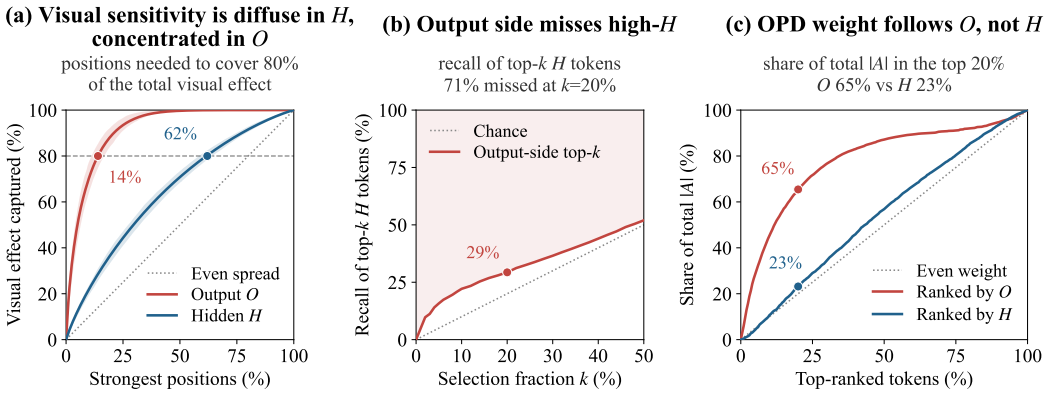


Figure 2: Motivation analysis of the representation–output mismatch. (a) Visual dependence is more diffuse in H than in O. (b) Output-side ranking misses high-H positions. (c) OPD weight follows O, not H.

position t, and let $\ell^T_{i,t}(I) = \log p_T(y_{i,t} \mid I, q_i, \mathbf{y}_{i,<t})$. We measure the corresponding representation- and output-level changes as

$$d^{H,\ell}_{i,t}(I') = \frac{\left\|\mathbf{h}^{(\ell)}_{i,t}(I_i) - \mathbf{h}^{(\ell)}_{i,t}(I')\right\|_2}{\max\left(\left\|\mathbf{h}^{(\ell)}_{i,t}(I_i)\right\|_2, \epsilon\right)}, \qquad d^{O}_{i,t}(I') = \left|\ell^T_{i,t}(I_i) - \ell^T_{i,t}(I')\right|. \tag{4}$$

We remove the change induced by the JPEG control and define

$$H_{i,t} = \frac{1}{L}\sum_{\ell=1}^{L}\left[d^{H,\ell}_{i,t}(I^{\text{shuf}}_i) - d^{H,\ell}_{i,t}(I^{\text{jpg}}_i)\right]_+, \qquad O_{i,t} = \left[d^{O}_{i,t}(I^{\text{shuf}}_i) - d^{O}_{i,t}(I^{\text{jpg}}_i)\right]_+. \tag{5}$$

Here, $H_{i,t}$ measures representation-level visual sensitivity, whereas $O_{i,t}$ measures how strongly the same perturbation is exposed through the probability of the sampled token. Figure 2(a) shows that the strongest 14% of positions ranked by O account for 80% of the output-side visual effect, whereas covering 80% of the representation-side effect requires 62% of the positions ranked by H, indicating that visual dependence is substantially more diffuse inside the teacher.

A local geometric view helps explain why representation-level changes need not be equally visible at the sampled-token output. Consider the teacher representation immediately before the language-model head at a fixed response position, let $\Delta\mathbf{h}_{i,t}$ denote its image-induced displacement, and define the sampled-token output-sensitive direction at the unperturbed representation as

$$\mathbf{g}_{i,t} = \nabla_{\mathbf{h}} \log p_T(y_{i,t} \mid \mathbf{h})\big|_{\mathbf{h}=\mathbf{h}_{i,t}(I_i)}. \tag{6}$$

A first-order expansion around $\mathbf{h}_{i,t}(I_i)$ gives

$$\Delta \log p_T(y_{i,t}) = \mathbf{g}^\top_{i,t}\Delta\mathbf{h}_{i,t} + O\big(\|\Delta\mathbf{h}_{i,t}\|_2^2\big). \tag{7}$$

Thus, even a substantial representation displacement can remain weakly exposed at the sampled-token output when it is poorly aligned with the corresponding output-sensitive direction. Appendix A.1 provides the full derivation and extends this view beyond a single output direction.

The mismatch is also evident when the two signals localize response positions. The output-side top 20% recovers only 29% of the top-20% representation-sensitive positions, leaving 71% unselected (Figure 2(b)), and their within-response rank association is weak ($\rho \approx 0.13$). Moreover, the top 20% positions ranked by O receive 65% of the total standard-OPD coefficient mass $|A|$, whereas the top 20% ranked by H receive only 23%, close to the 20% share expected under an unrelated ranking (Figure 2(c)). Positions in the top 10% by H and bottom 50% by O constitute about 5.0% of valid positions but receive only 0.81% of the total $|A|$ mass.

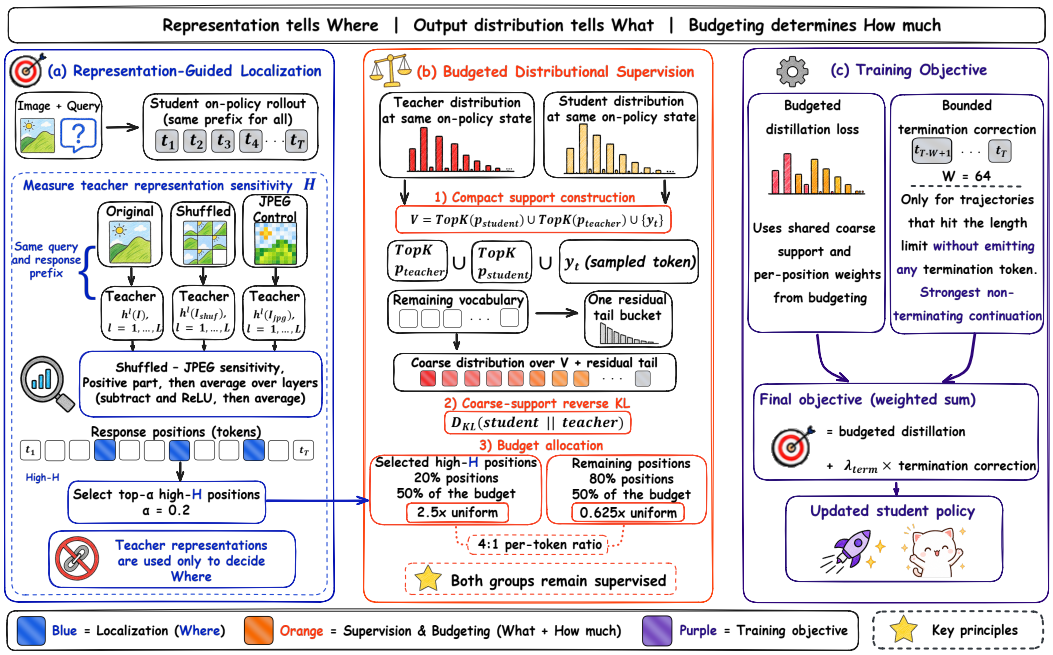


Figure 3: Overview of RGB-OPD. (a) Teacher representation sensitivity localizes response positions for additional supervision without direct hidden-state alignment. (b) A compact teacher–student output support provides budgeted distributional supervision while retaining supervision on all valid positions of eligible responses. (c) The budgeted objective is combined with a bounded termination correction for truncated trajectories.

Taken together, teacher representations expose visually sensitive positions that are only partially captured by output-side signals and receive little additional emphasis under standard OPD. This motivates using representation sensitivity to determine where output-space supervision should be concentrated.

3.3 Representation-Guided Budgeting for On-Policy Distillation

Motivated by the representation–output mismatch in Section 3.2, we propose Representation-Guided Budgeting for On-Policy Distillation (RGB-OPD). RGB-OPD uses teacher representation sensitivity to determine where output-space supervision should be concentrated, while the teacher output distribution specifies what the student should match and a budgeting rule controls how much supervision is assigned to different response positions. Figure 3 summarizes the framework.

3.3.1 Representation-Guided Localization

For each eligible response with $n_i = \sum_{t=1}^{T_i} m_{i,t} \geq T_{\min} = 64$, we select the top $\alpha = 0.2$ fraction of valid positions ranked by $H_{i,t}$:

$$\mathcal{S}_i = \mathrm{Top}_{\lceil \alpha n_i \rceil}(\{H_{i,t} \mid m_{i,t} = 1\}), \qquad \bar{\mathcal{S}}_i = \{t \mid m_{i,t} = 1\} \setminus \mathcal{S}_i. \tag{8}$$

We use $T_{\min} = 64$ to avoid applying the within-response budgeting rule to very short trajectories, where the selected–remaining split is more sensitive to integer rounding; these responses constitute only about 3% of the generated samples in our primary setting. Computing $H_{i,t}$ requires teacher evaluation under the original, shuffled, and JPEG-controlled image views for the same fixed student-generated prefix; these teacher-side computations are outside the gradient path. The exact finite-length clipping rule that guarantees both groups are nonempty is given in Appendix B.4.

At this selection fraction, the output-side ranking in Figure 2(b) recovers only 29% of the most representation-sensitive positions. The resulting sets $\mathcal{S}_i$ and $\bar{\mathcal{S}}_i$ define the two groups used by the budgeting rule below.

3.3.2 BUDGETED DISTRIBUTIONAL SUPERVISION

Localization identifies where additional supervision is needed, but it does not determine what the student should match. We therefore use a compact distributional target that exposes a broader portion of the teacher–student output discrepancy at each on-policy state (Zaidi et al., 2025; Fu et al., 2026).

Let $p_{\theta,i,t}$ and $p_{T,i,t}$ denote the student and teacher output distributions. We construct

$$\mathcal{V}_{i,t} = \mathrm{TopK}(p_{\theta,i,t}) \cup \mathrm{TopK}(p_{T,i,t}) \cup \{y_{i,t}\}, \tag{9}$$

where $K = 100$. The support retains dominant modes of both student and teacher and explicitly includes the sampled token. Probability mass outside $\mathcal{V}_{i,t}$ is aggregated into a single residual bucket, yielding coarse distributions $\widetilde{p}_{\theta,i,t}$ and $\widetilde{p}_{T,i,t}$; the exact construction is given in Appendix B.4. The token-level discrepancy is then measured by reverse KL:

$$D_{i,t} = D_{\mathrm{KL}}(\widetilde{p}_{\theta,i,t} \| \widetilde{p}_{T,i,t}) \,. \tag{10}$$

Unlike the sampled-token OPD objective in Eq. 3, RGB-OPD directly backpropagates through this analytical distribution-matching objective at the student-visited states. The likelihood-ratio form in Eq. 2 is therefore not required for the coarse reverse-KL term; the on-policy property is retained through the student-generated trajectories on which the teacher distributions are queried.

This distributional view complements representation-guided localization. While the sampled-token signal probes only one output-sensitive direction at each state, the compact support exposes discrepancies across multiple high-probability teacher and student modes. It therefore provides a richer output-space target at positions identified by representation sensitivity. Appendix A.1 formalizes this multi-direction view. The coarse-support construction preserves discrepancies over dominant output modes without requiring full-vocabulary distillation; its exact KL decomposition, equality condition, and support-refinement property are given in Appendix A.2.

We allocate a fixed response-level supervision budget between the representation-sensitive and remaining positions:

$$\mathcal{L}_i^{\mathrm{bud}} = \beta \frac{1}{|\mathcal{S}_i|} \sum_{t \in \mathcal{S}_i} D_{i,t} + (1-\beta) \frac{1}{|\bar{\mathcal{S}}_i|} \sum_{t \in \bar{\mathcal{S}}_i} D_{i,t}. \tag{11}$$

Unlike hard position selection, Eq. 11 retains supervision over all valid positions while reallocating a larger share of the response-level budget toward the representation-sensitive subset. Let $a_i = |\mathcal{S}_i|/n_i$ denote the realized selected fraction. The relative per-token supervision between selected and remaining positions is

$$\Gamma(a_i, \beta) = \frac{\beta(1-a_i)}{(1-\beta)a_i}. \tag{12}$$

We use $\beta = 0.5$; with the nominal $a_i = 0.2$, the selected 20% receive 50% of the response-level budget, giving a $4{:}1$ per-token supervision ratio.

The same reallocation has a simple gradient interpretation. Let $g_{\mathcal{S}}$ and $g_{\bar{\mathcal{S}}}$ denote the mean token gradients over the selected and remaining positions. The token-uniform and budgeted response-level gradients are $g_{\mathrm{uni}} = a_i g_{\mathcal{S}} + (1-a_i) g_{\bar{\mathcal{S}}}$ and $g_{\mathrm{bud}} = \beta g_{\mathcal{S}} + (1-\beta) g_{\bar{\mathcal{S}}}$, respectively, so

$$g_{\mathrm{bud}} - g_{\mathrm{uni}} = (\beta - a_i)\,(g_{\mathcal{S}} - g_{\bar{\mathcal{S}}}) \,. \tag{13}$$

Thus, under a fixed response-level budget, the update changes precisely through the gradient contrast between representation-sensitive and remaining positions. Appendix A.3 gives the full derivation, the expected behavior under random selection, and additional interpretations of the budgeting rule.

3.3.3 BOUNDED TERMINATION CORRECTION

Long on-policy reasoning trajectories introduce a separate failure mode: a response may reach the maximum generation length without emitting a valid termination token, in which case it is truncated and counted as incorrect under our evaluation protocol. We therefore apply a correction only to the final W valid positions of such trajectories.

Let $\mathcal{Z}$ denote the set of termination-token IDs, let $e^{\dagger} \in \mathcal{Z}$ be the designated termination token, and let $v_{i,t}^{\star} = \arg\max_{v \notin \mathcal{Z}} p_{\theta,i,t}(v)$ denote the strongest non-terminating continuation at each corrected

tail position. We define the per-position termination hinge as

$$\ell_{i,t}^{\mathrm{term}} = \left[\delta + \log p_{\theta,i,t}(v_{i,t}^{\star}) - \log p_{\theta,i,t}(e^{\dagger})\right]_{+}. \tag{14}$$

The termination loss $\mathcal{L}^{\mathrm{term}}$ averages this hinge over all gated tail positions of truncated trajectories in the minibatch. The hinge becomes inactive once the designated termination token satisfies the prescribed margin. We use $W = 64$ and $\delta = 2.0$; exact gating and aggregation, compact-support recovery of $v_{i,t}^{\star}$, and additional properties are provided in Appendix B.4 and Appendix A.4. This correction is applied only to trajectories that reach the fixed generation limit, providing a bounded stabilization mechanism for this failure mode.

3.3.4 Training Objective

Averaging $\mathcal{L}_i^{\mathrm{bud}}$ over responses with $n_i \geq T_{\min}$ gives $\mathcal{L}^{\mathrm{bud}}$. The final RGB-OPD objective is

$$\mathcal{L}_{\text{RGB-OPD}} = \mathcal{L}^{\mathrm{bud}} + \lambda_{\mathrm{term}}\mathcal{L}^{\mathrm{term}}, \tag{15}$$

where $\lambda_{\mathrm{term}} = 0.010$. The two terms address complementary aspects of training: $\mathcal{L}^{\mathrm{bud}}$ reallocates distributional supervision according to representation sensitivity, whereas $\mathcal{L}^{\mathrm{term}}$ is activated only for truncated trajectories to correct a separate termination failure mode. Gradients flow only through the student output probabilities in the coarse reverse-KL objective and termination correction; teacher probabilities, representations, sensitivity scores, selected sets, and discrete support membership are treated as fixed supervision quantities.

4 Experiments

4.1 Experimental Setup

Models and training. We consider two Qwen3-VL teacher–student settings (Bai et al., 2025): 8B→2B as the primary setting and 32B→8B for scaling. All trained methods use the same fixed 2,096-example Geometry3K training subset (Lu et al., 2021), with the vision encoder frozen. For the main comparison, we train the 8B→2B models for 5 epochs (655 optimization steps) and the 32B→8B models for 1 epoch (131 steps). All training runs are conducted on 8 NVIDIA H200 GPUs with 141 GB memory per GPU. All main-table experiments use random seed 42 unless otherwise stated. RGB-OPD uses $\alpha = 0.2$ and $\beta = 0.5$.[1] Additional implementation details, evaluation protocols, hyperparameter sensitivity, and robustness analyses are provided in Appendix B.

Baselines and evaluation. We compare with the pretrained Base student, chain-of-thought supervised fine-tuning (CoT-SFT), Standard OPD (Agarwal et al., 2024), and OPRD-Bridge (Yang et al., 2026), our adaptation of OPRD to the current multimodal on-policy training pipeline. Standard OPD serves as our primary teacher–student distillation baseline. We evaluate out-of-domain transfer on WeMath (Qiao et al., 2025), MathVerse (Zhang et al., 2024), MMBench (Liu et al., 2024), MMStar (Chen et al., 2024), and VisOnlyQA (Kamoi et al., 2024). Geometry3K is used only for training. We use greedy decoding and strict-match accuracy; responses that reach the maximum generation length without a valid termination token or fail to yield a parseable answer are counted as incorrect. **Avg-5** is the macro-average over the five evaluation benchmarks.

4.2 Main Results

Table 1 reports the main results. In the Qwen3-VL-8B→2B setting, RGB-OPD raises Avg-5 from 49.98 to 50.78 over Standard OPD, with higher point estimates on all five benchmarks and the largest gains on MathVerse (+2.03) and WeMath (+1.50). RGB-OPD also outperforms OPRD-Bridge on all five benchmarks, improving Avg-5 from 50.18 to 50.78. At the larger 32B→8B scale, Avg-5 increases from 65.52 to 66.43, again with higher scores on all five benchmarks, including +2.54 on MathVerse, +0.96 on VisOnlyQA, and +0.92 on WeMath. Overall, RGB-OPD yields higher point estimates than Standard OPD across all ten benchmark–scale comparisons.

To assess robustness across random seeds, we additionally evaluate Standard OPD and RGB-OPD over five matched seeds in the 8B→2B setting. RGB-OPD improves Avg-5 from 50.02 ± 0.18

[1] Anonymized repository with the core implementation; full code will be released upon acceptance.

Table 1: Main results across two Qwen3-VL teacher–student scales. Scores are greedy strict-match accuracy (%); Avg-5 is the macro-average over the five out-of-domain benchmarks.

	Mathematical Reasoning		General Multimodal Reasoning			
Method	WeMath	MathVerse	MMBench	MMStar	VisOnlyQA	**Avg-5**
Setting I: Qwen3-VL-8B → 2B, 5 epochs (655 steps)						
Base	51.72	20.30	82.31	55.27	35.83	49.09
CoT-SFT	51.90	22.34	81.38	54.80	33.91	48.87
Standard OPD	50.63	23.86	83.07	56.53	35.83	49.98
OPRD-Bridge	51.21	25.76	82.37	56.07	35.48	50.18
RGB-OPD (ours)	**52.13**	**25.89**	**83.16**	**56.80**	**35.91**	**50.78**
Δ vs. Standard OPD	**+1.50**	**+2.03**	**+0.09**	**+0.27**	**+0.08**	**+0.80**
Setting II: Qwen3-VL-32B → 8B, 1 epoch (131 steps)						
Base	71.78	39.85	88.73	68.33	53.83	64.50
CoT-SFT	74.14	40.61	89.54	69.00	53.57	65.37
Standard OPD	73.91	40.10	90.14	69.00	54.43	65.52
OPRD-Bridge	74.71	40.74	**90.32**	68.53	**55.39**	65.94
RGB-OPD (ours)	**74.83**	**42.64**	90.16	**69.13**	**55.39**	**66.43**
Δ vs. Standard OPD	**+0.92**	**+2.54**	**+0.02**	**+0.13**	**+0.96**	**+0.91**

to 50.64 $\pm$ 0.31, with a paired improvement of +0.62 $\pm$ 0.40 points. Full per-benchmark means, standard deviations, and paired differences are reported in Appendix B.5.

4.3 Ablation and Controlled Analysis

Selector and component ablations. Figure 4(a) compares budget-matched selector controls and leave-one-out ablations in the 8B→2B setting. Random and Sampled-O use the same coarse objective, termination correction, selection fraction, and supervision budget as RGB-OPD, differing only in how selected positions are identified. The w/o H-Budget and w/o Term variants remove representation-guided reallocation and the termination correction, respectively.

RGB-OPD achieves the largest gain on MathVerse (+2.03), exceeding Random and Sampled-O (+1.14 each) and w/o H-Budget (+1.40). On MMStar, Random, Sampled-O, and w/o H-Budget all degrade performance, whereas RGB-OPD improves accuracy by +0.27. These results suggest that non-uniform or sampled-output-guided supervision alone is not consistently beneficial, while representation-guided budgeting yields positive gains on both tasks. Removing the termination correction lowers MathVerse accuracy by 2.28 percentage points but leaves MMStar unchanged, indicating a substantial but task-dependent effect under our fixed maximum-length evaluation protocol.

Robustness to the construction of H. We further vary the block-shuffle granularity used to construct H while keeping all other training settings fixed. Across 2×2, 4×4, and 8×8 grids, RGB-OPD consistently outperforms Standard OPD on both MathVerse (24.87–26.40 vs. 23.86) and MMStar (56.80–57.47 vs. 56.53). These results suggest that the benefit of representation-guided budgeting is not tied to the specific 4×4 partition used by the default estimator. Full results are reported in Appendix B.8.

Discrepancy at representation-sensitive positions. We evaluate all checkpoints on a fixed diagnostic set of RGB-OPD response prefixes and measure the coarse-support reverse KL D_H at the top-20% positions ranked by H. Figure 4(b) shows that RGB-OPD yields the largest reduction relative to Standard OPD on both MathVerse (3.11%) and MMStar (0.65%). This shared-prefix evaluation isolates checkpoint differences at identical response states; it does not evaluate discrepancy under each model's own on-policy state distribution.

Effect of the termination correction. As shown in Figure 4(c), the termination correction reduces maximum-length truncation on four of five benchmarks, lowering the macro-average rate from 23.69% to 21.88% and the micro-average from 17.38% to 16.15%. The fraction of responses

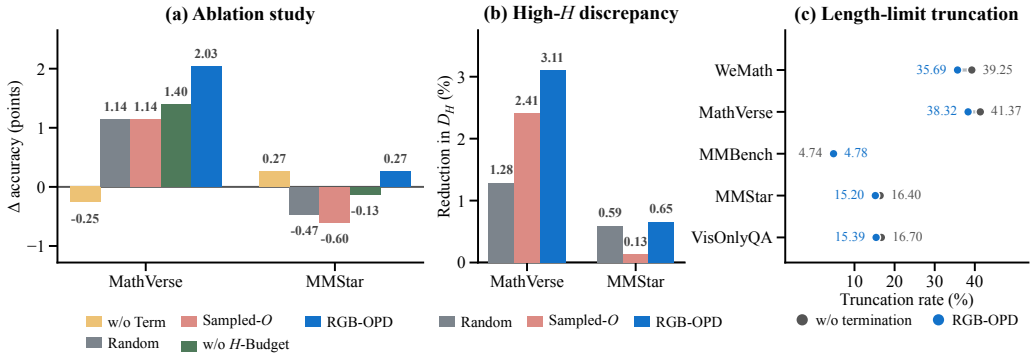


Figure 4: Controlled analyses in the Qwen3-VL-8B→2B setting. (a) Accuracy changes relative to Standard OPD for selector controls and leave-one-out ablations. (b) Relative reduction in coarse teacher–student discrepancy D_H at the top-20% H-ranked positions on shared RGB-OPD response states. (c) Length-limit truncation rates with and without the bounded termination correction.

Table 2: Training efficiency and wall-clock-matched comparison in the Qwen3-VL-8B→2B setting.

Method	Steps	Time/Step	Peak Memory	Avg. Response Length
Standard OPD	655	33.27 s (1.00×)	68.21 GB (1.00×)	2,662.3
RGB-OPD	655	51.19 s (1.54×)	71.43 GB (1.05×)	2,521.8 (−5.28%)

Method	WeMath	MathVerse	MMBench	MMStar	VisOnlyQA	Avg-5
Standard OPD (655 steps)	50.63	23.86	83.07	56.53	35.83	49.98
Standard OPD (1,008 steps)	52.01	23.35	83.11	55.27	**36.43**	50.04
RGB-OPD (655 steps)	**52.13**	**25.89**	**83.16**	**56.80**	35.91	**50.78**

shorter than 64 tokens changes only from 2.98% to 3.12%, suggesting that the reduction in truncation is not driven by substantially earlier stopping.

4.4 Training Efficiency

We profile Standard OPD and RGB-OPD in the Qwen3-VL-8B→2B setting on 8 NVIDIA H200 GPUs, averaging 50 steady-state optimization steps after 10 warm-up steps. Table 2 reports the training efficiency and wall-clock-matched comparison. The additional cost of RGB-OPD mainly arises from teacher-side representation probing and compact-support computation. Its average response length decreases from 2,662.3 to 2,521.8 tokens, so the higher cost is not due to longer generations. When Standard OPD is extended to 1,008 steps to approximately match the wall-clock cost of RGB-OPD at 655 steps, it reaches an Avg-5 of 50.04, still 0.74 points below RGB-OPD. The overhead is confined to training and introduces no additional inference-time computation.

5 Conclusion

RGB-OPD uses teacher representation sensitivity to localize where output-space supervision should be concentrated in multimodal on-policy distillation. Under a shared distributional objective and supervision budget, representation-guided localization yields positive gains over matched random and sampled-token output-guided selectors on both analyzed tasks and achieves the lowest coarse teacher–student discrepancy at representation-sensitive positions among the compared checkpoints. The bounded termination correction addresses a complementary failure mode by reducing maximum-length truncation on four of five out-of-domain benchmarks. Across two Qwen3-VL teacher–student scales, RGB-OPD improves Standard OPD, supporting teacher representation sensitivity as an effective localization signal for budgeted output-space distillation.

AI USE STATEMENT

We used generative AI tools for language editing during the preparation of this manuscript. All AI-assisted edits were critically reviewed and verified by the authors, who take full responsibility for the technical claims, experimental results, and final presentation of the paper.

ETHICS STATEMENT

This work studies model-training methodology using existing multimodal reasoning benchmarks and pretrained vision-language models. We do not collect new human-subject data or introduce new personally identifiable information. Our experiments are conducted for research purposes and do not involve deployment in high-stakes decision-making settings. As with other model distillation and post-training techniques, the proposed method may inherit limitations or biases from the teacher model, student model, and training data. We therefore encourage careful evaluation before applying the method beyond the research settings considered in this work.

REPRODUCIBILITY STATEMENT

We provide the complete RGB-OPD training objective and methodological details in Section 3. Experimental settings and evaluation protocols are described in Section 4.1, with detailed training hyperparameters, dataset splits and sample counts, representation-sensitivity computation, optimization details, and evaluation procedures provided in Appendix B. The theoretical properties of the representation–output analysis, coarse distributional objective, budgeting rule, and bounded termination correction are provided in Appendix A. We report the random seeds, hardware configuration, optimization steps, and decoding protocol used in the experiments to facilitate reproduction of the reported results.

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*, volume 2024, pp. 21246–21263, 2024.

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.

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. Vold: Reasoning transfer from llms to vision-language models via on-policy distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26209–26218, 2026.

Davide Caffagni, Alberto Compagnoni, Federico Melis, Sara Sarto, Pier Luigi Dovesi, Mark Granroth-Wilding, Marcella Cornia, and Lorenzo Baraldi. Mind the heads: Topological representation alignment for multimodal llms. *arXiv preprint arXiv:2606.23885*, 2026.

Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, Zhucun Xue, Yong Liu, and Xiang Bai. Llava-kd: A framework of distilling multimodal large language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 239–249. IEEE, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Ruoxi Cheng, Haoxuan Ma, Zhengfei Hai, Yiyan Huang, Ranjie Duan, Tianle Zhang, Xu Yang, Ziyi Ye, and Xingjun Ma. Adversarial orthogonal disentanglement for lvlm hallucination mitigation. *arXiv preprint arXiv:2605.25377*, 2026.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Ryo Kamoi, Yusen Zhang, Sarkar Snigdha Sarathi Das, Ranran Haoran Zhang, and Rui Zhang. Visonlyqa: Large vision language models still struggle with visual perception of geometric information. *arXiv preprint arXiv:2412.00947*, 2024.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv preprint arXiv:2603.11137*, 2026.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6121–6133, 2020.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, et al. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026.

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

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6774–6786, 2021.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20023–20070, 2025.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Jingchen Sun, Shaobo Han, Deep Patel, Wataru Kohno, Can Jin, and Changyou Chen. Uncertainty-aware knowledge distillation for multimodal large language models. *arXiv preprint arXiv:2603.21426*, 2026.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 4323–4332, 2019.

Almog Tavor, Itay Ebenspanger, Neil Cnaan, and Mor Geva. Rethinking selective knowledge distillation. *arXiv preprint arXiv:2602.01395*, 2026.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation. *arXiv preprint arXiv:2606.01249*, 2026.

Shilin Xu, Xiangtai Li, Haobo Yuan, Lu Qi, Yunhai Tong, and Ming-Hsuan Yang. Llavadi: What matters for multimodal large language models distillation. *arXiv preprint arXiv:2407.19409*, 2024.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026.

Shenzhi Yang, Guangcheng Zhu, Bowen Song, Haobo Wang, Mingxuan Xia, Xing Zheng, Yingfan Ma, Zhongqi Chen, Weiqiang Wang, Junbo Zhao, et al. Oprd: On-policy representation distillation. *arXiv preprint arXiv:2606.06021*, 2026.

Hee Suk Yoon, Eunseop Yoon, Jaehyun Jang, SooHwan Eom, Ji Woo Hong, Mark Hasegawa-Johnson, Qi Dai, Chong Luo, and Chang D Yoo. Decomposed on-policy distillation for vision-language reasoning: Steering gradients for visual grounding. *arXiv preprint arXiv:2606.00564*, 2026.

Heeji Yoon, Jaewoo Jung, Junwan Kim, Hyungyu Choi, Heeseong Shin, Sangbeom Lim, Honggyu An, Chaehyun Kim, Jisang Han, Donghyun Kim, et al. Visual representation alignment for multimodal large language models. *arXiv preprint arXiv:2509.07979*, 2025.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Mohd Abbas Zaidi, Akhil Kedia, Jinwoo Ahn, Taehwak Kwon, Kangwook Lee, Haejun Lee, Joohyung Lee, et al. Sparse logit sampling: Accelerating knowledge distillation in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18085–18108, 2025.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

A THEORETICAL ANALYSIS OF RGB-OPD

This appendix provides additional theoretical analysis for RGB-OPD. We study the local representation–output geometry underlying output attenuation, the information retained by coarse-support distillation, the gradient effect of representation-guided budgeting, and the local properties of the bounded termination correction.

All derivations condition on fixed student-generated prefixes. Unless stated otherwise, teacher quantities, representation-guided selection sets, and discrete vocabulary support sets are treated as constants under differentiation; gradients do not propagate through teacher computations, ranking or top-K membership, or token selection. Probability identities are stated under exact arithmetic.

A.1 REPRESENTATION-TO-OUTPUT ATTENUATION

Local representation-to-output geometry. Section 3.2 uses the local projection in Eq. 7 to motivate the representation–output mismatch. Here we derive this relation explicitly and examine how the same representation displacement can be observed through one or multiple output-sensitive directions.

We focus on the final response representation immediately before the teacher's language-model head, after any final normalization. Let $\mathbf{h} \in \mathbb{R}^d$ denote this representation at a fixed on-policy response position under the original image I, and let $\Delta\mathbf{h} = \mathbf{h}(I') - \mathbf{h}(I)$ be the displacement induced by an image perturbation $I \to I'$. For vocabulary token v, let the corresponding logit be $a_v(\mathbf{h}) = \mathbf{w}_v^\top \mathbf{h} + b_v$, with next-token distribution

$$p_T(v \mid \mathbf{h}) = \frac{\exp(a_v(\mathbf{h}))}{\sum_{v'} \exp(a_{v'}(\mathbf{h}))}.$$

The output-sensitive direction is

$$\mathbf{g}_v = \nabla_{\mathbf{h}} \log p_T(v \mid \mathbf{h}) = \mathbf{w}_v - \sum_{v'} p_T(v' \mid \mathbf{h}) \mathbf{w}_{v'}. \tag{16}$$

Unless otherwise stated, $\mathbf{g}_v$ below is evaluated at the unperturbed representation $\mathbf{h}(I)$. A first-order Taylor expansion around $\mathbf{h}(I)$ gives

$$\Delta \log p_T(v) = \mathbf{g}_v^\top \Delta\mathbf{h} + \mathcal{O}\big(\|\Delta\mathbf{h}\|_2^2\big).$$

For the sampled token $v = y_{i,t}$ and the corresponding displacement $\Delta\mathbf{h}_{i,t}$, this recovers Eq. 7. Representation sensitivity is therefore associated with the magnitude of the image-induced displacement, whereas the change exposed through a particular output token depends on the projection of this displacement onto $\mathbf{g}_v$.

Writing $\Delta\mathbf{h} = \|\Delta\mathbf{h}\|_2 \mathbf{u}$ and $\mathbf{g}_v = \|\mathbf{g}_v\|_2 \mathbf{z}_v$ with $\|\mathbf{u}\|_2 = \|\mathbf{z}_v\|_2 = 1$, the first-order relation gives

$$|\Delta \log p_T(v)| \approx \|\Delta\mathbf{h}\|_2 \|\mathbf{g}_v\|_2 \left|\mathbf{z}_v^\top \mathbf{u}\right|.$$

Thus, a substantial representation displacement can remain weakly expressed at a particular output token when the displacement is poorly aligned with that token's output-sensitive direction.

Single-direction attenuation. To illustrate this mechanism, consider an idealized directional model. Conditioned on a fixed nonzero displacement $\Delta\mathbf{h}$, suppose that the normalized output-sensitive direction $\mathbf{z}_v$ is isotropically distributed on the unit sphere in $\mathbb{R}^d$. Here, $\mathbf{z}_v = \mathbf{g}_v / \|\mathbf{g}_v\|_2$ is defined for $\mathbf{g}_v \neq \mathbf{0}$, and $\mathbf{u} = \Delta\mathbf{h} / \|\Delta\mathbf{h}\|_2$. Then

$$\mathbb{E}\Big[\big(\mathbf{z}_v^\top \mathbf{u}\big)^2 \mid \mathbf{u}\Big] = \frac{1}{d}, \qquad \mathbb{E}\left[\frac{(\mathbf{g}_v^\top \Delta\mathbf{h})^2}{\|\mathbf{g}_v\|_2^2} \;\middle|\; \Delta\mathbf{h}\right] = \frac{\|\Delta\mathbf{h}\|_2^2}{d}. \tag{17}$$

The first identity follows directly from rotational symmetry: choosing coordinates with $\mathbf{u} = (1, 0, \ldots, 0)^\top$, all coordinates of an isotropic unit vector have identical second moments and satisfy $\sum_{j=1}^d \mathbb{E}[z_{v,j}^2] = 1$, so $\mathbb{E}[z_{v,1}^2] = 1/d$. The second identity follows by substituting $\mathbf{g}_v = \|\mathbf{g}_v\|_2 \mathbf{z}_v$ and $\Delta\mathbf{h} = \|\Delta\mathbf{h}\|_2 \mathbf{u}$.

This isotropic model is used only to formalize the attenuation mechanism and is not intended as a claim that actual vocabulary gradients or output embeddings are isotropic. In real language models, these directions may exhibit substantial anisotropy. The deterministic point remains that a large $\|\Delta\mathbf{h}\|_2$ does not imply a large projection onto any particular $\mathbf{g}_v$. The sampled-token output-side sensitivity probe therefore observes only one output-sensitive projection of the image-induced representation displacement.

Multi-token observability. Consider an output support $\mathcal{V} = \{v_1, \ldots, v_m\}$ and let $\mathbf{G}_\mathcal{V}$ stack the corresponding row vectors $\mathbf{g}_v^\top$. The first-order output displacement observed through this support is $\Delta\boldsymbol{\ell}_\mathcal{V} \approx \mathbf{G}_\mathcal{V}\Delta\mathbf{h}$. Define the aggregate first-order observable energy as

$$E_\mathcal{V}(\Delta\mathbf{h}) = \|\mathbf{G}_\mathcal{V}\Delta\mathbf{h}\|_2^2 = \sum_{v\in\mathcal{V}} \left(\mathbf{g}_v^\top \Delta\mathbf{h}\right)^2. \tag{18}$$

For two nested supports $\mathcal{V}_1 \subseteq \mathcal{V}_2$,

$$E_{\mathcal{V}_2}(\Delta\mathbf{h}) - E_{\mathcal{V}_1}(\Delta\mathbf{h}) = \sum_{v\in\mathcal{V}_2\setminus\mathcal{V}_1} \left(\mathbf{g}_v^\top \Delta\mathbf{h}\right)^2 \geq 0. \tag{19}$$

This monotonicity is deterministic and requires no isotropy assumption. It concerns aggregate observable energy and does not require newly added output gradients to introduce linearly independent directions.

For geometric interpretation, define the corresponding normalized-direction energy

$$\overline{E}_\mathcal{V}(\Delta\mathbf{h}) = \sum_{v\in\mathcal{V}} \left(\mathbf{z}_v^\top \Delta\mathbf{h}\right)^2.$$

Under the idealized isotropic model above,

$$\mathbb{E}\left[\overline{E}_\mathcal{V}(\Delta\mathbf{h}) \mid \Delta\mathbf{h}\right] = \frac{|\mathcal{V}|}{d}\|\Delta\mathbf{h}\|_2^2.$$

This follows from linearity of expectation and does not require the directions $\{\mathbf{z}_v\}_{v\in\mathcal{V}}$ to be mutually independent. Thus, observing a broader output support exposes a nondecreasing aggregate set of first-order output projections of the representation displacement.

The analysis in this subsection concerns local perturbations of the teacher's final response representation immediately before the language-model head. The multi-layer representation sensitivity H used by RGB-OPD, as well as the behavior under finite image perturbations, is characterized empirically in Section 3.2. The geometric analysis establishes that high representation sensitivity can coexist with weak sampled-token output sensitivity; it does not by itself imply that such positions have greater task value.

A.2 Coarse-Support Reverse-KL Decomposition

RGB-OPD avoids full-vocabulary distillation by explicitly retaining high-probability teacher and student tokens and aggregating the remaining vocabulary into a single residual bucket. Let Ω denote the complete vocabulary, let $\mathcal{V} \subseteq \Omega$ denote the explicit support, and let $\mathcal{R} = \Omega \setminus \mathcal{V}$. Let p and q denote the student and teacher distributions, respectively, at a fixed on-policy state, with $p(v) > 0$ and $q(v) > 0$ for all $v \in \Omega$. Define

$$p(\mathcal{R}) = \sum_{v\in\mathcal{R}} p(v), \qquad q(\mathcal{R}) = \sum_{v\in\mathcal{R}} q(v),$$

and the corresponding coarse distributions

$$\widetilde{p} = \left(\{p(v)\}_{v\in\mathcal{V}}, p(\mathcal{R})\right), \qquad \widetilde{q} = \left(\{q(v)\}_{v\in\mathcal{V}}, q(\mathcal{R})\right).$$

Proposition 1. For a nonempty residual set $\mathcal{R}$, the full-vocabulary reverse KL admits the exact decomposition

$$D_{\mathrm{KL}}(p\|q) = D_{\mathrm{KL}}(\widetilde{p}\|\widetilde{q}) + p(\mathcal{R})D_{\mathrm{KL}}\left(p(\cdot \mid \mathcal{R})\|q(\cdot \mid \mathcal{R})\right). \tag{20}$$

Consequently,

$$D_{\mathrm{KL}}(\widetilde{p}\|\widetilde{q}) \leq D_{\mathrm{KL}}(p\|q),$$

with equality when $\mathcal{R} = \varnothing$, or, for nonempty $\mathcal{R}$, if and only if $p(\cdot \mid \mathcal{R}) = q(\cdot \mid \mathcal{R})$.

Proof. Separating the full KL over $\mathcal{V}$ and $\mathcal{R}$ gives

$$D_{\mathrm{KL}}(p\|q) = \sum_{v\in\mathcal{V}} p(v)\log\frac{p(v)}{q(v)} + \sum_{v\in\mathcal{R}} p(v)\log\frac{p(v)}{q(v)}.$$

For $v \in \mathcal{R}$, write $p(v) = p(\mathcal{R})p(v \mid \mathcal{R})$ and $q(v) = q(\mathcal{R})q(v \mid \mathcal{R})$. The residual term becomes

$$p(\mathcal{R})\log\frac{p(\mathcal{R})}{q(\mathcal{R})} + p(\mathcal{R})D_{\mathrm{KL}}\left(p(\cdot \mid \mathcal{R})\|q(\cdot \mid \mathcal{R})\right).$$

The first term combines with the contributions on $\mathcal{V}$ to form $D_{\mathrm{KL}}(\widetilde{p}\|\widetilde{q})$, proving Eq. 20. Nonnegativity of KL gives the lower bound and the equality condition.

Thus, the coarse objective preserves all teacher–student discrepancies on the explicit support and the discrepancy in total residual mass, while discarding only redistribution internal to the aggregated residual bucket.

Support-refinement monotonicity. Let $\mathcal{V}_1 \subseteq \mathcal{V}_2 \subseteq \Omega$, and let $(\widetilde{p}_1, \widetilde{q}_1)$ and $(\widetilde{p}_2, \widetilde{q}_2)$ denote the corresponding coarse distributions. Because the partition induced by $\mathcal{V}_2$ refines that induced by $\mathcal{V}_1$, the KL chain rule yields

$$D_{\mathrm{KL}}(\widetilde{p}_1\|\widetilde{q}_1) \le D_{\mathrm{KL}}(\widetilde{p}_2\|\widetilde{q}_2) \le D_{\mathrm{KL}}(p\|q). \tag{21}$$

The first inequality arises from a nonnegative conditional KL term inside the residual bucket of the coarser partition.

These identities characterize objective values for fixed p, q, and a fixed explicit support $\mathcal{V}$. They do not by themselves bound the discrepancy between gradients of the coarse and full-vocabulary objectives. In particular, a small residual mass $p(\mathcal{R})$ alone is insufficient to guarantee a small objective-value gap without also controlling the conditional tail divergence $D_{\mathrm{KL}}(p(\cdot \mid \mathcal{R})\|q(\cdot \mid \mathcal{R}))$. They also do not imply that uniformly weighted coarse-support KL is equivalent to standard sampled-token OPD.

A.3 Budget Conservation and Selector-Dependent Updates

We next formalize how representation-guided budgeting changes the update induced by token-level distributional supervision. Consider a response containing T valid positions, let $\mathcal{S}$ denote the selected representation-sensitive subset, let $m = |\mathcal{S}|$, and define the realized selected fraction $a = m/T$ with $1 \le m < T$. We use a rather than the nominal selection fraction α because finite responses may differ slightly from α due to integer rounding or ties.

Let D_t denote the coarse-support reverse-KL discrepancy at position t. The budget coefficient applied to each position is

$$b_t = \begin{cases} \dfrac{\beta}{m}, & t \in \mathcal{S}, \\ \dfrac{1-\beta}{T-m}, & t \notin \mathcal{S}. \end{cases} \tag{22}$$

The response-level objective is $\mathcal{L}^{\mathrm{bud}} = \sum_{t=1}^{T} b_t D_t$. By construction,

$$\sum_{t=1}^{T} b_t = 1, \qquad \frac{b_{\mathrm{selected}}}{b_{\mathrm{remaining}}} = \frac{\beta(1-a)}{(1-\beta)a}.$$

When $\beta = a$, the coefficient reduces exactly to $1/T$ at every position. Budget conservation refers only to the sum of the scalar coefficients; it does not imply conservation of the norm or direction of the resulting parameter gradient.

Proposition 2 (Selector-dependent gradient shift). Define the mean token gradients within the selected and remaining groups as

$$\mathbf{g}_{\mathcal{S}} = \frac{1}{m}\sum_{t\in\mathcal{S}} \nabla_\theta D_t, \qquad \mathbf{g}_{\overline{\mathcal{S}}} = \frac{1}{T-m}\sum_{t\notin\mathcal{S}} \nabla_\theta D_t.$$

The corresponding budgeted and token-uniform response-level gradients are

$$\mathbf{g}_b = \beta \mathbf{g}_{\mathcal{S}} + (1-\beta)\mathbf{g}_{\overline{\mathcal{S}}}, \qquad \mathbf{g}_u = a\mathbf{g}_{\mathcal{S}} + (1-a)\mathbf{g}_{\overline{\mathcal{S}}}.$$

Subtracting the two expressions gives

$$\boxed{\mathbf{g}_b - \mathbf{g}_u = (\beta - a)\left(\mathbf{g}_{\mathcal{S}} - \mathbf{g}_{\overline{\mathcal{S}}}\right)}.$$

This is the generic-response form of Eq. 13. Thus, under a fixed response-level budget, the selector changes the update only through the difference between the mean gradients of the selected and remaining positions. When $\beta = a$, or when the two group-average gradients coincide, budgeting recovers the uniform update.

Proposition 3 (Expected update under size-matched random selection). Fix the model parameters, the student-generated prefixes, and the token-level discrepancies $\{D_t\}_{t=1}^{T}$. Let $\mathcal{S}_{\mathrm{rand}}$ be sampled uniformly from all subsets of $\{1, \ldots, T\}$ having size m, and apply the same group budget β to $\mathcal{S}_{\mathrm{rand}}$. Then

$$\mathbb{E}_{\mathcal{S}_{\mathrm{rand}}}\left[\mathbf{g}_b(\mathcal{S}_{\mathrm{rand}})\right] = \mathbf{g}_u. \tag{23}$$

Proof. Every position is included with probability $m/T = a$. Therefore, both the selected-group mean gradient and the complementary-group mean gradient have expectation

$$\frac{1}{T}\sum_{t=1}^{T} \nabla_\theta D_t = \mathbf{g}_u.$$

Taking their weighted sum with coefficients β and $1-\beta$ proves Eq. 23. This result concerns selector randomness at a fixed model state and fixed prefixes; it does not imply equivalence between random selection and uniform allocation over a complete training trajectory, where subsequent model states and generated prefixes may differ.

Proposition 4 (Conditional local advantage of informative selection). Let $F(\theta)$ denote the full-vocabulary distillation risk evaluated on the same fixed student-generated prefixes, and suppose that F has a locally Lipschitz gradient around θ. For sufficiently small $\tau > 0$,

$$\begin{aligned} &F(\theta - \tau\mathbf{g}_b) - F(\theta - \tau\mathbf{g}_u) \\ &\quad = -\tau(\beta - a)\left\langle \nabla F(\theta), \mathbf{g}_{\mathcal{S}} - \mathbf{g}_{\overline{\mathcal{S}}} \right\rangle + \mathcal{O}\left(\tau^2\left[\|\mathbf{g}_b\|_2^2 + \|\mathbf{g}_u\|_2^2\right]\right). \end{aligned} \tag{24}$$

Consequently, if

$$\beta > a, \qquad \left\langle \nabla F(\theta), \mathbf{g}_{\mathcal{S}} - \mathbf{g}_{\overline{\mathcal{S}}} \right\rangle > 0,$$

then, for sufficiently small τ, the budgeted update yields a lower post-update risk than the uniform update:

$$F(\theta - \tau\mathbf{g}_b) < F(\theta - \tau\mathbf{g}_u).$$

Proof. Local smoothness gives, for a fixed update direction $\mathbf{g}$,

$$F(\theta - \tau\mathbf{g}) = F(\theta) - \tau\left\langle \nabla F(\theta), \mathbf{g} \right\rangle + \mathcal{O}(\tau^2\|\mathbf{g}\|_2^2).$$

Applying this expansion to $\mathbf{g}_b$ and $\mathbf{g}_u$, subtracting the two expressions, and using Eq. 13 gives Eq. 24. Under the stated conditions, the first-order term is strictly negative and therefore dominates the second-order remainder for sufficiently small τ.

This proposition is conditional: the analysis does not establish that representation sensitivity H necessarily selects positions satisfying the gradient-alignment condition above. The controlled localization experiments provide indirect empirical evidence for the practical utility of the representation-guided selector, rather than a direct verification of the gradient-alignment condition in Proposition 4. The result is also a local gradient-descent statement on fixed prefixes and does not constitute a guarantee for full AdamW training dynamics or downstream accuracy.

Entropy-regularized interpretation. The two-level budget also admits an entropy-regularized allocation interpretation. Let $\mathbf{q} = (q_1, \ldots, q_T)$ denote a distribution over token-level distillation budget, let $u_t = 1/T$ denote uniform allocation, and let $s_t = \mathbb{1}[t \in \mathcal{S}]$. Consider

$$\max_{\mathbf{q} \in \Delta_T} \left[\eta \sum_{t=1}^{T} q_t s_t - D_{\mathrm{KL}}(\mathbf{q} \| \mathbf{u}) \right], \qquad \eta \geq 0. \tag{25}$$

The first term rewards budget placed on representation-sensitive positions, whereas the KL term regularizes the allocation toward uniform supervision. The unique solution is

$$q_t^\star = \frac{u_t \exp(\eta s_t)}{\sum_{j=1}^{T} u_j \exp(\eta s_j)}.$$

This follows from the stationarity condition of the corresponding Lagrangian; strict convexity of $D_{\mathrm{KL}}(\mathbf{q} \| \mathbf{u})$ for $u_t > 0$ makes the solution unique.

Since $|\mathcal{S}| = aT$ by definition of the realized fraction, the selected-group budget is

$$\beta = \frac{ae^\eta}{ae^\eta + (1-a)}.$$

Equivalently,

$$e^\eta = \frac{\beta(1-a)}{a(1-\beta)}, \qquad \Gamma(a, \beta) = \frac{\beta(1-a)}{(1-\beta)a} = e^\eta. \tag{26}$$

Because $\eta \geq 0$, this interpretation corresponds to $\beta \in [a, 1)$. When $\eta = 0$, we obtain $\beta = a$ and uniform token allocation. When $\eta > 0$, selected positions receive larger per-token supervision, and as $\eta \to \infty$, the allocation approaches hard concentration on the selected subset.

For the nominal RGB-OPD configuration $\alpha = 0.2$ and $\beta = 0.5$, a response with realized fraction $a = 0.2$ has $\Gamma = 4$ and $\eta = \log 4$. For finite responses affected by integer rounding or ties, the exact per-response concentration factor is instead $\Gamma(a, 0.5)$ with $a = m/T$.

A.4 Properties of the Bounded Termination Correction

We finally analyze the termination-margin objective introduced in Section 3.3.3 under the multiple-stop-token generation semantics used by the model. For notational simplicity, we suppress the response index i from position-wise quantities throughout this subsection; sequence-level quantities retain i where needed. Let $\mathcal{Z} \subset \Omega$ denote the set of all token IDs that terminate generation, and let $e^\dagger \in \mathcal{Z}$ denote the designated termination token whose probability is explicitly promoted by the correction. For a truncated response and a corrected tail position t, define the strongest non-terminating continuation as

$$v_t^\star = \arg\max_{v \notin \mathcal{Z}} p_\theta(v).$$

The termination hinge is

$$\ell_t^{\mathrm{term}} = \left[\delta + \log p_\theta(v_t^\star) - \log p_\theta(e^\dagger) \right]_+, \qquad \delta > 0. \tag{27}$$

The term "bounded" refers to the bounded logit-space gradient and the finite-margin saturation characterized below, not to the numerical value of the hinge loss itself, which is not upper bounded.

Exact recovery of the strongest continuation from the compact support. In implementation, the competitor is selected from the compact RGB-OPD support rather than by explicitly scanning the full vocabulary. The statement below applies at the parameter state at which the support is constructed, so that the support contains the current student top-K set being analyzed. If that support contains $\mathrm{TopK}(p_\theta)$ and

$$K > |\mathcal{Z}|,$$

then

$$\max_{v \in \mathrm{TopK}(p_\theta) \setminus \mathcal{Z}} p_\theta(v) = \max_{v \notin \mathcal{Z}} p_\theta(v). \tag{28}$$

Indeed, $\mathrm{TopK}(p_\theta)$ must contain at least one non-termination token when $K > |\mathcal{Z}|$. Moreover, a globally highest-probability non-termination token cannot be excluded from a valid top-K set

while a lower-probability token is included. Therefore, at the parameter state used to construct the support, selecting the strongest non-terminating competitor from the compact support is exact rather than approximate. If the support were instead frozen while the student parameters subsequently changed, the condition $K > |\mathcal{Z}|$ alone would no longer guarantee exact recovery for the updated distribution.

Proposition 5. If $\ell_t^{\mathrm{term}} = 0$, then

$$p_\theta(e^\dagger) \geq e^\delta \max_{v \notin \mathcal{Z}} p_\theta(v). \tag{29}$$

Consequently, when $\delta > 0$, every globally highest-probability token belongs to the termination set,

$$\arg\max_{v \in \Omega} p_\theta(v) \subseteq \mathcal{Z},$$

so greedy decoding selects a termination action at that position.

Proof. The condition $\ell_t^{\mathrm{term}} = 0$ implies

$$\log p_\theta(e^\dagger) - \log p_\theta(v_t^\star) \geq \delta.$$

Exponentiating and using the definition of $v_t^\star$ gives Eq. 29. Since $\delta > 0$, the designated termination token has strictly larger probability than every non-termination token. Another termination token may nevertheless have even higher probability than $e^\dagger$, so the guarantee is that every global maximizer is a termination token, not that $e^\dagger$ is necessarily the unique global maximizer.

Bounded logit-space gradient and margin saturation. Let $\mathbf{z}$ denote the student logits at the corrected position. Whenever the hinge in Eq. 27 is active and the strongest non-terminating competitor $v_t^\star$ is unique, the shared log-partition term cancels:

$$\log p_\theta(v_t^\star) - \log p_\theta(e^\dagger) = z_{v_t^\star} - z_{e^\dagger}.$$

Therefore,

$$\nabla_{\mathbf{z}} \ell_t^{\mathrm{term}} = \mathbf{e}_{v_t^\star} - \mathbf{e}_{e^\dagger}, \qquad \left\| \nabla_{\mathbf{z}} \ell_t^{\mathrm{term}} \right\|_2 = \sqrt{2}. \tag{30}$$

At ties between multiple strongest non-terminating competitors, the hinge is nondifferentiable, but any valid subgradient with respect to $\mathbf{z}$ has Euclidean norm at most $\sqrt{2}$. This is a bound in logit space; the corresponding parameter-space gradient additionally depends on the network Jacobian.

Whenever the termination margin is strictly satisfied,

$$\log p_\theta(e^\dagger) - \log p_\theta(v_t^\star) > \delta,$$

the hinge lies in the interior of its zero region and $\nabla_\theta \ell_t^{\mathrm{term}} = 0$ almost everywhere, excluding measure-zero boundaries associated with ties in the maximizing competitor or the hinge transition. The correction therefore does not continue increasing termination preference at that fixed prefix once the prescribed finite margin has been achieved.

For a truncated response i, let $\mathcal{E}_i$ denote its corrected tail positions and define

$$\mathcal{L}_i^{\mathrm{term}} = \frac{1}{|\mathcal{E}_i|} \sum_{t \in \mathcal{E}_i} \ell_t^{\mathrm{term}}.$$

Since each hinge term is nonnegative, $\mathcal{L}_i^{\mathrm{term}} = 0$ implies $\ell_t^{\mathrm{term}} = 0$ for every $t \in \mathcal{E}_i$, and therefore Eq. 29 holds at every corrected tail position.

Together, these properties clarify the role of the bounded termination correction. The hinge enforces a finite separation between termination and continuation, has a bounded logit-space gradient while active, and becomes inactive after the prescribed margin is strictly satisfied. These are local properties at the corrected prefixes. Because model parameters are shared across prefixes, the resulting update can affect termination behavior elsewhere; the analysis therefore does not guarantee that early termination cannot increase at other prefixes or over the course of full training.

B Implementation and Evaluation Details

This appendix records the training configuration, evaluation data, sensitivity computation, controlled selectors, and exact implementation details used in our experiments. Hyperparameters are fixed across downstream benchmarks unless stated otherwise.

B.1 Training Configuration

All methods are trained on the same fixed 2,096-example Geometry3K subset with the vision encoder frozen. We consider two Qwen3-VL teacher–student settings: 8B→2B as the primary setting and 32B→8B for scaling. For the main comparison, the former is trained for 5 epochs (655 optimization steps), while the latter is trained for 1 epoch (131 steps). Methods in the main comparison within the same setting share the student initialization, training data, and prescribed number of optimization steps. For OPRD-Bridge, we adapt the OPRD representation-distillation objective to the same Qwen3-VL multimodal on-policy pipeline while retaining the shared student initialization, training subset, rollout protocol, and prescribed optimization budget within each teacher–student setting. We additionally train Standard OPD for 1,008 steps as the wall-clock-matched control reported in Table 2.

We use AdamW with a learning rate of 1×10^{-6}. Each optimization step contains 16 prompts and 4 student rollouts per prompt, yielding 64 trajectories per step. Responses are capped at 4,096 tokens, and each rollout batch is used for one optimization pass. All reported experiments use random seed 42 unless otherwise stated. Training uses 8 NVIDIA H200 GPUs, and the maximum gradient norm for on-policy methods is 10. We do not tune the learning rate or other training hyperparameters on Geometry3K or any downstream evaluation benchmark.

Table 3 summarizes the training and RGB-OPD hyperparameters.

Table 3: Main training and RGB-OPD hyperparameters.

Hyperparameter	**Value**
Optimizer	AdamW
Learning rate	1×10^{-6}
Prompt batch size	16
Rollouts per prompt	4
Trajectories per step	64
Maximum response length	4,096
Optimization passes per rollout batch	1
Maximum gradient norm (on-policy)	10
Random seed (main runs)	42
Vision encoder	Frozen
Selection fraction α	0.2
Selected-group budget β	0.5
Compact-support Top-K	100
Minimum eligible response length $T_{\min}$	64
Termination tail length W	64
Termination margin δ	2.0
Termination weight λ_{term}	0.010

All RGB-OPD hyperparameters are fixed across the five downstream evaluation benchmarks.

B.2 Datasets and Benchmarks

Table 4 summarizes the training and evaluation data. Geometry3K (Lu et al., 2021) is used only for training, and none of the five downstream benchmarks is used for training or hyperparameter selection. Avg-5 is the unweighted macro-average over the five evaluation sets.

Table 4: Datasets used for training and out-of-domain evaluation. Sample counts correspond to the instances decoded by our evaluation pipeline.

Dataset	Split	# Samples	Primary capability
Training data			
Geometry3K	Training subset	2,096	Geometry reasoning
Out-of-domain evaluation			
WeMath	testmini	1,740	Visual mathematical reasoning
MathVerse	testmini, Vision-Only	788	Visual mathematical reasoning
MMBench	DEV EN	4,329	General multimodal reasoning
MMStar	Full benchmark	1,500	Vision-dependent multimodal reasoning
VisOnlyQA	VLMEvalKit evaluation split	1,150	Fine-grained visual perception

The counts in Table 4 refer to examples actually processed by our evaluation pipeline rather than the nominal size of each benchmark release. For MathVerse, we use the Vision-Only subset of `testmini`, resulting in 788 examples. Every method within a teacher–student setting is evaluated on the same instances with the same decoding and answer-processing pipeline.

B.3 Sensitivity Signals and Controlled Selectors

Representation-sensitivity probe. The representation-sensitivity signal H is computed from teacher forward passes with the textual query and student-generated response prefix held fixed. Each example is evaluated with three visual inputs: the original image I_i, a shuffled image I_i^{shuf} that disrupts global spatial structure, and a mildly JPEG-compressed image I_i^{jpg} used as a low-level perturbation control. Image dimensions and all textual tokens remain unchanged across the three passes.

To obtain I_i^{shuf}, the image is divided into a 4×4 grid and the resulting 16 blocks are permuted. The permutation is deterministic for each example: its seed is derived from the SHA-256 hash of ``42:sample_id'', with `sample_id` replaced by the example identifier. If the sampled permutation is the identity, we apply a cyclic rotation. This preserves local content inside each block while disrupting the global arrangement.

The JPEG control is encoded with quality 90, `subsampling=0` (4:4:4 chroma sampling), and `optimize=False`. It is intended to account for representation changes caused by low-level image processing rather than semantic disruption. Following Eq. 5, the JPEG-induced change is subtracted from the shuffled-image change before taking the positive part.

Hidden states are collected from all language-model transformer blocks, excluding the embedding layer. For a response containing T valid tokens, response-aligned states are extracted with the slice `[-T-1:-1]`, which matches the response prediction positions and excludes the final boundary state. The same alignment is used for the original, shuffled, and JPEG-controlled passes, and the corrected layer-wise sensitivities are averaged uniformly to obtain $H_{i,t}$.

Teacher forward passes, image perturbations, sensitivity scores, and ranking are outside the gradient path. The same definition of H is used in the motivation analysis and RGB-OPD training; teacher hidden states are used only to rank response positions and are never optimized as student alignment targets.

Controlled selectors. For the selector controls in Section 4.3, we keep the coarse distributional objective, selection fraction, group-level budget, and bounded termination correction identical to RGB-OPD, changing only how the selected positions are determined. For each eligible response, the number of selected positions follows the same finite-length rule as Eq. 8 and Appendix B.4.

Random selects the corresponding number of valid response positions uniformly at random. **Sampled-O** instead ranks valid positions by the sampled-token output sensitivity $O_{i,t}$ defined in Eq. 5 and selects the highest-ranked positions. Both controls use the same $\alpha = 0.2$, $\beta = 0.5$, compact-support reverse-KL objective, and termination correction as RGB-OPD. Thus, their only

difference from RGB-OPD is the signal used to determine where the additional supervision budget is concentrated.

B.4 Exact RGB-OPD Implementation

Finite-length position selection. For an eligible response with $n_i \geq T_{\min} = 64$, the number of selected positions is

$$k_i = \min\{\max(\lceil \alpha n_i \rceil, 1), n_i - 1\}. \tag{31}$$

The k_i valid positions with the largest $H_{i,t}$ form $\mathcal{S}_i$, with the remaining valid positions forming $\bar{\mathcal{S}}_i$. This clipping rule guarantees that both groups are nonempty. Because k_i is integer-valued, the realized selected fraction $a_i = k_i/n_i$ can differ slightly from the nominal $\alpha = 0.2$.

Residual-bucket construction. Using the compact support $\mathcal{V}_{i,t}$ defined in Eq. 9, probability mass outside the explicit support is retained in a single residual bucket. For $M \in \{\theta, T\}$, we define

$$p^{\text{tail}}_{M,i,t} = 1 - \sum_{v \in \mathcal{V}_{i,t}} p_{M,i,t}(v), \tag{32}$$

and form

$$\widetilde{p}_{M,i,t} = \left(\{p_{M,i,t}(v)\}_{v \in \mathcal{V}_{i,t}}, p^{\text{tail}}_{M,i,t}\right). \tag{33}$$

The resulting student and teacher coarse distributions are compared using the reverse KL in Eq. 10. Appendix A.2 gives the corresponding exact decomposition and support-refinement properties.

Budgeted-loss aggregation. For each eligible response, Eq. 11 is evaluated using the realized sets $\mathcal{S}_i$ and $\bar{\mathcal{S}}_i$. Finite-length rounding therefore affects the exact number of positions in each group, while the selected- and remaining-group budgets remain β and $1 - \beta$, respectively. The minibatch budgeted loss averages $\mathcal{L}^{\text{bud}}_i$ uniformly over responses satisfying $n_i \geq T_{\min}$ and is set to zero if no response in the minibatch is eligible.

Bounded termination correction. The termination correction is activated only when a trajectory reaches the maximum response length without emitting a valid termination token. The termination set $\mathcal{Z}$ contains the model-defined <|im_end|> and <|endoftext|> tokens, with token IDs 151645 and 151643. We use <|im_end|> ($e^\dagger = 151645$) as the designated token and apply the correction to the final $W = 64$ valid response positions.

At each corrected position, $v^\star_{i,t}$ is the highest-probability student token outside $\mathcal{Z}$. It can be recovered exactly from the compact support at the same student parameter state: $\mathcal{V}_{i,t}$ contains the student Top-K set, and $K = 100 > |\mathcal{Z}|$, so at least one globally highest-probability non-termination token is present. Appendix A.4 gives the corresponding exact-recovery argument.

The hinge is defined in Eq. 14 with $\delta = 2.0$. Although it is written in terms of log probabilities, the shared log-partition cancels in $\log p_{\theta,i,t}(v^\star_{i,t}) - \log p_{\theta,i,t}(e^\dagger)$, so the margin is equivalently defined in logit space.

Let $\mathcal{I}_{\text{trunc}}$ denote trajectories that hit the generation limit without a valid termination token, and let $\mathcal{E}_i$ contain the final W valid positions of trajectory i. The gated minibatch positions are

$$\mathcal{G}_{\text{term}} = \{(i,t) \mid i \in \mathcal{I}_{\text{trunc}},\ t \in \mathcal{E}_i\}. \tag{34}$$

The termination objective averages the hinge over this pooled set:

$$\mathcal{L}^{\text{term}} = \frac{1}{|\mathcal{G}_{\text{term}}|} \sum_{(i,t) \in \mathcal{G}_{\text{term}}} \ell^{\text{term}}_{i,t}, \tag{35}$$

with $\mathcal{L}^{\text{term}} = 0$ if $\mathcal{G}_{\text{term}}$ is empty. Thus, gated positions are weighted uniformly across the minibatch rather than first averaging within each trajectory. We use $\lambda_{\text{term}} = 0.010$ in the final objective.

The hinge stops contributing once the designated termination token satisfies the prescribed margin against the strongest non-terminating continuation. Appendix A.4 analyzes its finite-margin saturation and bounded logit-space gradient.

B.5 MULTI-SEED EVALUATION

To assess robustness across random seeds, we repeat Standard OPD and RGB-OPD in the Qwen3-VL-8B→2B setting over five matched seeds (40, 41, 42, 43, 44). All runs use the same training data, optimization schedule, decoding protocol, and evaluation pipeline as the main experiments. Table 5 reports the mean and standard deviation across seeds.

RGB-OPD improves Avg-5 from 50.02 ± 0.18 to 50.64 ± 0.31. The paired Avg-5 difference is +0.62 ± 0.40 points, with positive mean paired improvements on all five evaluation benchmarks, although individual benchmark effects exhibit larger variability.

Table 5: Five-seed evaluation in the Qwen3-VL-8B→2B setting. Results are reported as mean ± standard deviation over five matched seeds. The difference row is computed per seed as RGB-OPD minus Standard OPD and then averaged.

Method	WeMath	MathVerse	MMBench	MMStar	VisOnlyQA	Avg-5
Standard OPD	50.59±0.38	24.59±0.70	82.85±0.37	55.72±0.77	36.37±0.36	50.02±0.18
RGB-OPD	**51.33±1.23**	**25.15±0.86**	**83.15±0.23**	**56.99±0.58**	**36.59±1.09**	**50.64±0.31**
Paired difference	**+0.75±1.52**	**+0.56±1.30**	**+0.30±0.30**	**+1.27±1.11**	**+0.23±0.86**	**+0.62±0.40**

B.6 ANALYSIS AND EVALUATION PROTOCOL

Motivation-analysis statistics. Figure 2 is computed from two frozen diagnostic probes, and all rankings are taken over valid response positions only.

Panel (a). Panel (a) uses $N_a = 192$ responses sampled with seed 42 from three multimodal reasoning datasets, with 64 responses from each dataset. The responses are generated once by a fixed early-stage SFT checkpoint used only for this diagnostic probe and distinct from the CoT-SFT baseline in the main experiments, under greedy decoding with a 128-token cap. Their mean response length is 98.5 tokens, and 85/192 responses reach the cap. The representation sensitivity H and sampled-token output effect O are both measured by the 8B teacher on the same frozen response token IDs.

For $s \in \{H, O\}$, let $\mathcal{T}_i^s(k)$ denote the top-k fraction of valid positions in response i after sorting by $s_{i,t}$. Signal concentration is measured by

$$C_s(k) = \frac{1}{N_a} \sum_{i=1}^{N_a} \frac{\sum_{t \in \mathcal{T}_i^s(k)} s_{i,t}}{\sum_{t:m_{i,t}=1} s_{i,t}}. \tag{36}$$

Panels (b) and (c). Panels (b) and (c) use $N_{bc} = 64$ prompts from the Geometry3K validation split. Response prefixes are frozen from the Standard OPD checkpoint at step 655 and truncated at 512 response tokens, yielding 30,386 valid response tokens. The same frozen prefixes are used for the H ranks, the O ranks, and the Standard OPD coefficient mass $|A|$, so all quantities in panels (b) and (c) are aligned token by token rather than evaluated on separate on-policy rollouts.

For the ranking-overlap and OPD-weight analyses, $H_{i,t}$ and $O_{i,t}$ are converted to within-response percentile ranks $r_{i,t}^H$ and $r_{i,t}^O$ before pooling valid positions. For $s \in \{H, O\}$, define

$$\mathcal{T}^s(k) = \left\{(i,t) \;\middle|\; m_{i,t} = 1,\ r_{i,t}^s \geq 1 - k\right\}. \tag{37}$$

The recall of representation-sensitive positions under output-side selection is

$$R_{O \to H}(k) = \frac{|\mathcal{T}^O(k) \cap \mathcal{T}^H(k)|}{|\mathcal{T}^H(k)|}, \tag{38}$$

whose expectation under independent rankings is k. The corresponding share of Standard OPD coefficient mass is

$$W_s(k) = \frac{\sum_{(i,t) \in \mathcal{T}^s(k)} |A_{i,t}|}{\sum_{i=1}^{N_{bc}} \sum_{t:m_{i,t}=1} |A_{i,t}|}. \tag{39}$$

The high-H/low-O diagnostic region in Section 3.2 is

$$\mathcal{G} = \left\{(i,t) \;\middle|\; r^{H}_{i,t} \geq 0.9,\ r^{O}_{i,t} \leq 0.5\right\}. \tag{40}$$

The 4×4 block shuffle and JPEG quality-90 control used in both probes are the same perturbations used to compute representation sensitivity during RGB-OPD training.

Evaluation protocol. Final checkpoints are evaluated with greedy decoding on WeMath, MathVerse, MMBench, MMStar, and VisOnlyQA. Geometry3K is used only for training. Within each teacher–student setting, every method uses the same benchmark instances, maximum response length, decoding procedure, and answer-processing pipeline.

We report strict-match accuracy and do not use an external model judge. A response is marked incorrect if it reaches the generation limit without a valid termination token or if the evaluation pipeline cannot extract a parseable final answer. Otherwise, the extracted answer is scored with the same benchmark-specific normalization and matching rule for every method. Avg-5 is the unweighted macro-average of the five out-of-domain accuracies.

Main-table results use the final checkpoint from each prescribed run. The 8B→2B setting uses 655 optimization steps and the 32B→8B setting uses 131 steps. RGB-OPD keeps α, β, K, $T_{\min}$, and the termination hyperparameters fixed across downstream benchmarks; no benchmark-specific tuning is performed.

B.7 Hyperparameter Sensitivity

We examine the sensitivity of RGB-OPD to the selection fraction α and selected-group budget β in the Qwen3-VL-8B→2B setting. All variants are trained for 655 steps and evaluated with greedy decoding on MathVerse and MMStar. We vary one hyperparameter at a time around the default setting $(\alpha, \beta) = (0.2, 0.5)$.

Table 6: Sensitivity of RGB-OPD to the selection fraction α and selected-group budget β.

α	β	MathVerse	MMStar
0.10	0.50	25.25	56.00
0.20	**0.50**	25.89	56.80
0.30	0.50	24.75	**56.93**
0.20	0.40	**26.40**	56.60
0.20	0.60	25.51	55.60

Performance remains competitive around the default setting, although the preferred allocation varies across tasks. Increasing α from 0.2 to 0.3 slightly improves MMStar but lowers MathVerse, while decreasing β to 0.4 improves MathVerse with a small decrease on MMStar. We therefore keep $(\alpha, \beta) = (0.2, 0.5)$ fixed across all downstream benchmarks rather than tuning the allocation separately for each task.

B.8 Robustness to the Construction of H

We evaluate the robustness of RGB-OPD to the spatial granularity of the block-shuffle perturbation used to construct H. Only the shuffle grid size is varied, while all other training settings are kept fixed. The 4×4 grid is the default used throughout the main experiments; the 2×2 and 8×8 variants are evaluated only as robustness checks and are not used for model selection.

Table 7: Robustness to the block-shuffle granularity used to construct H in the Qwen3-VL-8B→2B setting.

Shuffle grid	MathVerse	MMStar
Standard OPD	23.86	56.53
2×2	**26.40**	**57.47**
4×4 (default)	<u>25.89</u>	56.80
8×8	24.87	56.80

All three grid granularities outperform Standard OPD on both benchmarks. Although performance varies moderately with the shuffle granularity, the gains persist across the tested settings, indicating that RGB-OPD does not rely on the specific 4×4 partition used in the default configuration.

C Extended Related Work

This section complements the discussion in Section 2 by reviewing recent work on selective and adaptive distillation, distribution-level supervision, and representation-based knowledge transfer. We focus on works that are particularly relevant to the two design choices of RGB-OPD: identifying where supervision should be emphasized and determining what output-space information should be matched.

Selective and adaptive distillation. Recent work has questioned the conventional practice of assigning comparable distillation supervision to all tokens, classes, or samples. Rethinking Selective Knowledge Distillation systematically decomposes selective distillation along the position, vocabulary-class, and sample axes, and shows that student uncertainty can provide an effective signal for selecting informative positions (Tavor et al., 2026). TIP studies token importance specifically under on-policy distillation and identifies two complementary sources of useful supervision: positions with high student uncertainty and positions where a confident student substantially disagrees with the teacher (Xu et al., 2026). REOPOLD interprets OPD through a policy-optimization perspective and incorporates entropy-based token-level dynamic sampling together with reward clipping to stabilize training (Ko et al., 2026). TrOPD instead focuses on the reliability of teacher supervision under teacher–student distribution mismatch, restricting or modifying OPD updates according to agreement and trust-region criteria (Xing et al., 2026). Beyond OPD, uncertainty-aware multimodal distillation such as Beta-KD adaptively controls the strength of teacher supervision according to teacher uncertainty (Sun et al., 2026). These approaches establish that non-uniform supervision can improve either efficiency or optimization quality, but their selection signals are derived primarily from student uncertainty, teacher–student disagreement, or supervision reliability. RGB-OPD is complementary: its selection criterion is based on image-induced changes in the teacher's internal representations and is designed to expose visually dependent positions that may receive weak emphasis from output-side signals.

Distribution-level and sparse-output distillation. A second line of work studies how much of the teacher output distribution should be exposed to the student. DistiLLM investigates effective divergence objectives for large-language-model distillation on student-generated samples and highlights the importance of distribution-level teacher information beyond hard targets (Ko et al., 2024). More directly related to OPD, Revisiting On-Policy Distillation identifies the fragility of reducing teacher–student distribution matching to the sampled token and proposes matching a teacher top-K local support through truncated reverse KL (Fu et al., 2026). Sparse Logit Sampling studies a related computational problem in offline knowledge distillation, showing that naively retaining only the largest teacher logits can yield a biased approximation of the full teacher distribution and developing sampling-based alternatives for sparse-logit supervision (Zaidi et al., 2025). These studies motivate exposing a broader portion of the teacher distribution while controlling computational cost. In RGB-OPD, the compact support serves as the output-space target, while localization is determined independently by representation sensitivity. Our support combines dominant modes from both the teacher and student together with the sampled token and aggregates the remaining probability mass into a residual bucket; representation-guided budgeting is then applied over the resulting

coarse distribution. Thus, the central question in RGB-OPD is not merely which vocabulary entries to retain, but where distributional supervision should be emphasized based on visual dependence revealed inside the teacher.

Multimodal and representation-based knowledge transfer. Knowledge distillation for multimodal large language models has explored supervision at multiple representational levels. LLaVA-KD transfers both output-level knowledge and internal visual representations from larger multimodal models to smaller students (Cai et al., 2025), while LLAVADI systematically examines the roles of token- and logit-level alignment in multimodal model distillation (Xu et al., 2024). Other approaches focus more explicitly on internal representation structure: VIRAL aligns multimodal representations with pretrained visual foundation models (Yoon et al., 2025), and HeRA performs topology-aware alignment over selected attention heads (Caffagni et al., 2026). Representation geometry can also be used without conventional teacher–student feature matching; for example, AOD identifies hallucination-related latent directions in LVLM representations and intervenes on the resulting internal structure (Cheng et al., 2026). Under on-policy training, OPRD directly aligns teacher and student hidden states along trajectories generated by the student (Yang et al., 2026). RGB-OPD uses image-induced teacher representation sensitivity to rank response positions, while distillation is performed over output distributions. This separates representation-based localization from representation-matching objectives such as OPRD.