BEYOND OUTPUT ALIGNMENT: CAPABILITY DECOMPOSED ON-POLICY LATENT DISTILLATION FOR VISION-LANGUAGE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) has achieved strong results in LLMs by providing dense teacher supervision on student-visited states, reducing the train–inference mismatch. However, OPD supervision is mediated through next-token distributions: it teaches the student how the teacher would act, but does not explicitly transfer the intermediate representations supporting visual grounding and reasoning. Distilling the internal visual-reasoning states is promising, yet it faces two critical challenges: teacher and student representations may fall into different spaces, and their reasoning behaviors can differ substantially in wording, length, and granularity. Addressing these challenges, we introduce CD-OPLD, a capability-decomposed framework for latent distillation in VLMs. CD-OPLD starts from a structured cold-start checkpoint, then learns capability-specific visual and reasoning projectors to align student and teacher representations. During on-policy training, student-generated trajectories are paired with answer-correct teacher reference trajectories, and a monotonic semantic block matcher enables direct latent alignment between semantically corresponding visual and reasoning states. Distilling Qwen3.5-9B into Qwen3.5-2B, CD-OPLD consistently improves over a cold-start-matched OPD baseline across seven multimodal benchmarks. Code is available at `https://anonymous.4open.science/r/CD-OPLD-889F/`.

1 INTRODUCTION

Vision-language models (VLMs) achieve strong performance on visual question answering and multimodal reasoning, but their computational and memory costs motivate distillation into compact students. On-policy distillation (OPD) provides dense teacher supervision on student-generated prefixes, reducing the mismatch between training contexts and those encountered during inference Agarwal et al. (2024). However, its effectiveness in language models does not automatically carry over to vision-language models. Multimodal reasoning requires identifying task-relevant visual evidence, modeling relations among visual entities, and conducting visual-grounded reasoning. In our analysis of 200 stratified examples where an OPD-trained student answers incorrectly but the teacher succeeds, visual perception & recognition errors and problem understanding & conceptual knowledge errors account for 54.50% and 29.00% of total failures respectively (Figure 2). OPD constrains what the student should predict, but does not explicitly constrain what the student should internally represent about the image or the evolving reasoning state. Recent multimodal OPD methods strengthen visually relevant supervision, but their objectives remain primarily in the output space Liu et al. (2026); Yoon et al. (2026); Zhang et al. (2026a); Liang et al. (2026).

Can we directly supervise visual-grounding and reasoning capability alignment? Capability-aware distillation methods, including CompoDistill, LaViT, and OPLD, move toward this goal by explicitly transferring internal visual or multimodal reasoning signals Kim et al. (2026); Wu et al. (2026); Zhu et al. (2026). However, most of these works assume the teacher and student share the same latent representation space. Furthermore, existing works typically follow the vanilla OPD setting: teacher provides supervisions conditioned on student's potentially flawed prefix. A successful teacher trajectory exposes internal visual and reasoning states associated with a complete solution path, and provides richer signals for student to learn the evidence grounding and reasoning. However, these

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.

advances leave open how to establish meaningful correspondences between student and teacher trajectories that differ in wording, length, and reasoning granularity. Our key insight is to combine explicit capability-specific latent alignment with cross-trajectory semantic matching. We structure both rollouts into visual grounding, reasoning, and final-answer spans, pairing the student's on-policy trajectory with a correct teacher reference. The former captures student-visited states, while the latter supplies reliable visual-grounding and reasoning targets. We project student representations into the corresponding teacher latent spaces and align semantically matched visual and reasoning blocks, providing direct capability-level supervision without requiring token-wise correspondence or identical trajectory structures.

Based on this insight, we propose CD-OPLD, a three-stage framework comprising structured cold-start SFT, contrastive pretraining of capability-specific projectors, and joint on-policy output and latent distillation. The framework complements token-level OPD with explicit visual and reasoning representation alignment, supervising both what the student predicts and the internal representations supporting those predictions.

Our contributions are summarized as follows:

- We identify a capability-state gap in multimodal OPD: image-conditioned output supervision does not explicitly align the internal representations supporting visual grounding and reasoning.
- We introduce cross-trajectory semantic block matching to transfer visual and reasoning representations from a correct teacher reference to a student's on-policy rollout despite differences in wording and reasoning granularity.
- We develop CD-OPLD, a three-stage framework that bridges heterogeneous student-teacher latent spaces and combines capability-specific latent supervision with output-level OPD.
- Across seven benchmarks, CD-OPLD consistently outperforms standard off-policy and on-policy knowledge distillation methods. Compared with recent visual-aware distillation baselines, CD-OPLD leads performance on six of seven benchmarks. Among 200 stratified cases where pure OPD fails but the teacher succeeds, CD-OPLD corrects 40% of the errors.

2 Related Work

Knowledge Distillation and On-Policy Distillation. Knowledge distillation (KD) conventionally transfers knowledge from a high-capacity teacher to a compact student by matching softened output distributions Hinton et al. (2015). Autoregressive variants train on teacher-generated sequences Kim & Rush (2016) or optimize token-level divergences to retain richer predictive information Gu et al. (2024). Because such off-policy methods learn from fixed teacher trajectories, their training contexts may differ from those encountered during student generation. Generalized Knowledge Distillation (GKD) reduces this mismatch by querying the teacher on student-generated prefixes and distilling at states visited by the current student policy Agarwal et al. (2024). CD-OPLD follows this on-policy principle while extending supervision from next token distributions to capability-specific multimodal representations.

On-Policy Distillation for Vision-Language Models. Recent work adapts OPD to vision-language models by strengthening the visual signal within output-space supervision. VA-OPD emphasizes visually informative tokens and trajectories Liu et al. (2026); Decomposed OPD separates language-prior and visual-grounding gradients Yoon et al. (2026); and VAD reconstructs targets from visually attributable teacher evidence Zhang et al. (2026a). Visual Contrastive Self-Distillation further introduces contrastive supervision for visual grounding Liang et al. (2026). These methods refine where or how strongly output-level distillation is applied, but the transferred target remains primarily a next token distribution. In contrast, CD-OPLD retains OPD on student visited prefixes while directly aligning teacher visual grounding and reasoning states with semantically corresponding parts of the student's trajectory.

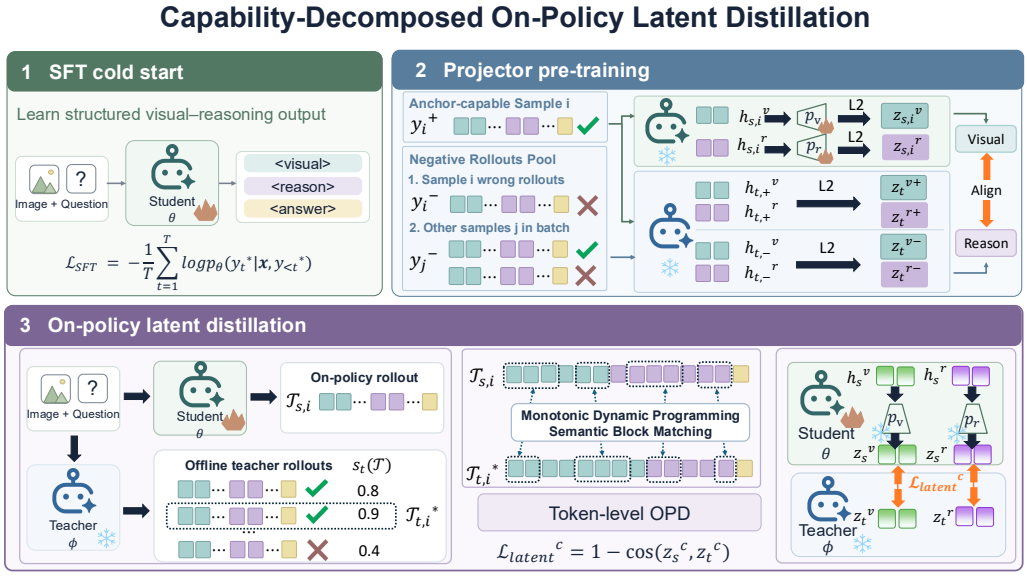

Figure 1: Overview of the proposed three-stage capability-decomposed on-policy latent distillation framework.

Representation and Latent Distillation. Representation distillation exposes intermediate knowledge that may be obscured by the output head, including layer-wise features Romero et al. (2015) and Transformer attention relations Wang et al. (2020). For multimodal reasoning, CompoDistill aligns visual attention Kim et al. (2026), whereas LaViT transfers latent visual thoughts derived from multimodal traces Wu et al. (2026). OPRD further aligns intermediate representations on shared on-policy rollouts Yang et al. (2026), and OPLD distills multimodal reasoning into latent states Zhu et al. (2026). Existing alignment, however, often assumes shared token positions, identical trajectories, or predefined latent units—correspondences that become brittle when student and teacher trajectories differ in wording, length, and reasoning granularity. CD-OPLD instead uses capability-specific projection and monotonic many-to-many semantic block matching to distill visual and reasoning representations across nonidentical trajectories without token-wise correspondence.

3 METHOD

To address the capability-state gap in multimodal on-policy distillation, we propose a three-stage framework that combines structured cold-start supervised fine-tuning (SFT) (section 3.1), capability-specific projector pretraining (section 3.2), and joint output-level OPD with capability-decomposed latent alignment (section 3.3). The overall architecture is illustrated in Figure 1.

3.1 STAGE 1: SFT COLD-START

To equip the student VLM with basic adherence to the required visual-reasoning output format, we perform structured cold-start SFT using complete, answer-correct teacher rollouts. Each rollout places visual-perception content inside `<visual></visual>`, reasoning content inside `<reason></reason>`, and the final answer inside `<final answer></final answer>`. A complete format-valid rollout example can be found in appendix A. We train on a stratified 4K subset of ViRL39K Wang et al. (2026) that is disjoint from the data used for subsequent on-policy distillation; its construction is detailed in Appendix C. The student minimizes the autoregressive loss over all T tokens of each gold rollout:

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta(y_t^* \mid \mathbf{x}, y_{<t}^*)\,, \tag{1}$$

3.2 Stage 2: Pre-train Projector: Align Student and Teacher Latent Spaces

To realize the effective latent distillation, we first need to align student's and teacher's latent space. We achieve this goal via pre-training two capability-specific projectors at the student side via unsupervised contrastive learning. We adopt the InfoNCE objective Chen et al. (2020); Oord et al. (2018) with the same split of the 8K subset of ViRL39K Wang et al. (2026) as later on-policy distillation. For each sample in the subset, we ask the teacher model to randomly sample 5 rollouts with temperature 1.0. We only keep samples with complete three-span structured rollouts as training data for this stage. Details of the completeness distribution of teacher offline rollouts are provided in Appendix D. To obtain more reliable visual-grounding and reasoning capability alignment, we restrict anchor-capable samples to those containing at least one correct-answer rollout. Specifically, we randomly sample a valid batch with N samples where valid means the number of anchor-capable samples is larger than zero: $B_a > 0$ and there are at least two different samples to form negative pairs.

For each correct rollout, we teacher-force the same image, question, prompt, and structured response through both models, extract token-level predictive hidden states from comparable late-middle layers (student layer 20/24 and teacher layer 28/32), and mean-pool them separately over the visual and reasoning spans Jiang et al. (2025). Two independent student-side projectors, P_v and P_r, implemented as lightweight MLPs, map the corresponding pooled student representations into the teacher's representation dimension. After L2 normalization, each projected student representation forms an anchor $z^c_{s,i,j}$, and the teacher representation from the same rollout forms its positive $z^{c,+}_{t,i,j}$, where $c \in \{v, r\}$. The negative set $\mathcal{N}_c(i)$ comprises teacher representations extracted in the same manner from (1) certified-wrong rollouts of the same sample i and (2) all complete rollouts of other samples in the batch. The resulting capability-specific contrastive loss is

$$\ell^c_{i,j} = -\log \frac{\exp\big(\mathrm{sim}\big(z^c_{s,i,j}, z^{c,+}_{t,i,j}\big)/\tau\big)}{\exp\big(\mathrm{sim}\big(z^c_{s,i,j}, z^{c,+}_{t,i,j}\big)/\tau\big) + \sum_{z \in \mathcal{N}_c(i)} \exp\big(\mathrm{sim}\big(z^c_{s,i,j}, z\big)/\tau\big)}, \qquad c \in \{v, r\} \tag{2}$$

where $z^{c,+}_{t,i,j}$ denotes positive part of the anchor extracted from the teacher model, $\mathcal{N}_c(i)$ denotes the negative set of current anchor. Since all the hidden representations are L2 normalized, the $\mathrm{sim}(z_i, z_j)$ equals to cosine similarity between the two vectors: $cos(z_i, z_j)$. τ denotes the temperature parameter. We denote the number of correct rollouts in an anchor-capable sample i as m_i, for a batch with B_a anchor-capable samples, we aggregate the capability-specific projector loss function as follows:

$$\mathcal{L}^c_{\text{Contrastive}} = \frac{1}{B_a} \sum_{i=1}^{B_a} \frac{1}{m_i} \sum_{j=1}^{m_i} \ell^c_{i,j}, \qquad c \in \{v, r\} \tag{3}$$

Finally, we train the two projectors together with the following objective:

$$\mathcal{L}_{\text{proj}} = \mathcal{L}^v_{\text{Contrastive}} + \mathcal{L}^r_{\text{Contrastive}} \tag{4}$$

3.3 Stage 3: On Policy Latent Distillation

Stage 3 combines token-level OPD on student-generated trajectories with capability-specific latent distillation from a selected correct teacher reference. We describe student rollout generation and teacher-reference selection in Section 3.3.1, semantic block matching in Section 3.3.2, matched-block visual and reasoning distillation in Section 3.3.3, and the final joint objective in Section 3.3.4.

3.3.1 Student On-Policy Rollout and Teacher Reference Selection

For each input $\mathbf{x}_i = (I_i, q_i)$, we sample K trajectories directly from the current student policy with prompt P as shown in appendix A:

$$\hat{\mathcal{T}}^k_{s,i} \sim \pi_\theta(\cdot \mid \mathbf{x}_i, P), \qquad k = 1, \ldots, K. \tag{5}$$

Among the pre-generated teacher offline rollouts, we retain the format-valid and answer-correct candidates $\mathcal{C}_i$ and select the highest-scoring one $\mathcal{T}^*_{t,i}$ as the teacher reference as follows:

$$\begin{aligned} \mathcal{C}_i &= \left\{ \mathcal{T}^j_{t,i} \mid A^j_{t,i} = A^*_i \text{ and } \mathcal{T}^j_{t,i} \text{ is format-valid} \right\}, \\ s_t\left(\mathcal{T}^j_{t,i}\right) &= \frac{1}{M_i} \sum_{m=1}^{M_i} \log p_\phi\left(a^*_{i,m} \mid \mathbf{x}_i, P, V\left(\mathcal{T}^j_{t,i}\right), R\left(\mathcal{T}^j_{t,i}\right), a^*_{i,<m}\right), \\ \mathcal{T}^*_{t,i} &= \arg\max_{\mathcal{T} \in \mathcal{C}_i} s_t(\mathcal{T}). \end{aligned} \tag{6}$$

Here, P is the input prompt; $V(\mathcal{T}^j_{t,i})$ and $R(\mathcal{T}^j_{t,i})$ are the visual and reason spans of the j-th teacher rollout; and $a^*_{i,m}$ is the m-th token of the ground-truth answer. To compute $s_t\left(\mathcal{T}^j_{t,i}\right)$, we replace the generated final answer of $\mathcal{T}^j_{t,i}$ with the ground-truth answer and teacher-force the resulting complete trajectory through the frozen teacher with parameter ϕ.

If no format-valid and answer-correct teacher rollout exists, i.e., $\mathcal{C}_i = \emptyset$, the sample does not participate in latent distillation. Its training objective automatically reduces to standard token-wise OPD.

3.3.2 Cross-Trajectory Semantic Block Matching

The student on-policy rollout and teacher reference may express corresponding content with different wording, length, and reasoning granularity. Token-wise alignment can therefore pair unrelated states, whereas whole-span mean pooling can dilute localized information such as decisive visual anchors and key mathematical derivations. We instead extract their visual and reasoning spans, exclude the final-answer span, and segment each span into an ordered sequence of semantic chunks.

We teacher-force both trajectories through the frozen teacher to obtain token representations in a common space. Consecutive chunks are grouped into candidate block pairs with varying numbers of student and teacher chunks, enabling many-to-many alignment. For each candidate $b = (b_s, b_t)$, we combine semantic, visual-anchor, and equation-feature similarities:

$$s_b = \begin{cases} 0, & \text{if an explicit semantic conflict is detected,} \\ \dfrac{\sum_{f \in \mathcal{F}_b} w_f s_f(b)}{\sum_{f \in \mathcal{F}_b} w_f}, & \text{otherwise,} \end{cases} \tag{7}$$

where $\mathcal{F}_b \subseteq \{h, a, e\}$ contains the available features. The semantic score is $s_h(b) = \max\{0, \cos(\bar{h}^{teacher}_{b_s}, \bar{h}^{teacher}_{b_t})\}$, while the visual-anchor and equation scores are the Jaccard similarities $s_a(b) = J(\mathcal{A}(b_s), \mathcal{A}(b_t))$ and $s_e(b) = J(\mathcal{E}(b_s), \mathcal{E}(b_t))$. Explicitly conflicting candidates receive zero score.

We then use monotonic dynamic programming Gale & Church (1993) to select $\mathcal{M}^*$, the globally optimal non-crossing sequence of matched block pairs, while allowing unmatched chunks on either side to be skipped:

$$\mathcal{M}^* = \arg\max_{\substack{\mathcal{M} \text{ monotonic, non-overlapping} \\ s_b \geq \delta \ \ \forall b \in \mathcal{M}}} \left[\sum_{b \in \mathcal{M}} s_b - \gamma_s N^s_{\text{gap}} - \gamma_t N^t_{\text{gap}} - \eta \sum_{b \in \mathcal{M}} (|b_s| + |b_t| - 2) \right], \tag{8}$$

where δ filters low-similarity candidates, the gap terms penalize excessive skipping, and the group-size term discourages unnecessary chunk merging. Visual and reasoning spans are aligned separately. Parsing, chunking, feature extraction, conflict filtering, and malformed-response handling are detailed in Appendix E; hyperparameter values are reported in Table 6.

3.3.3 Visual & Reason Latent Distillation

For each capability $c \in \{v, r\}$, let $\mathcal{M}^c_{i,k}$ denote the matched block pairs between the k-th student rollout of sample i and its teacher reference. For each $(b_s, b_t) \in \mathcal{M}^c_{i,k}$, we teacher-force the student on its on-policy rollout and the frozen teacher on the selected reference, then extract the predictive hidden states corresponding to the tokens in the two blocks from student layer 20 and teacher layer

28 respectively. The student side representations are mean-pooled and further go through the frozen capability-specific projectors. We then normalize the block representations as follows:

$$z^c_{s,i,k}(b_s) = \mathrm{Norm}\left(P_{s,c}\left(\frac{1}{|\Omega_s(b_s)|}\sum_{u\in\Omega_s(b_s)} h_{s,i,k,u}\right)\right),$$
$$z^c_{t,i}(b_t) = \mathrm{Norm}\left(\frac{1}{|\Omega_t(b_t)|}\sum_{u\in\Omega_t(b_t)} h_{t,i,u}\right), \qquad c \in \{v, r\} \tag{9}$$

where $\Omega_s(b_s)$ and $\Omega_t(b_t)$ are the predictive token positions of the matched student and teacher blocks, respectively, and $\mathrm{Norm}(z)$ denotes the L2-normalized representation of z.

We compute the visual and reason latent losses with the same objective:

$$\mathcal{L}^c_{\text{latent}} = \frac{1}{|\mathcal{V}_c|}\sum_{(i,k)\in\mathcal{V}_c}\frac{1}{|\mathcal{M}^c_{i,k}|}\sum_{(b_s,b_t)\in\mathcal{M}^c_{i,k}}\left[1-\cos\left(z^c_{s,i,k}(b_s), z^c_{t,i}(b_t)\right)\right], \qquad c\in\{v,r\} \tag{10}$$

where $\mathcal{V}_c$ contains the student rollouts with at least one matched block pair for capability c. The teacher and both projectors remain frozen, so the latent losses update only the student model.

3.3.4 Final On-Policy Latent Distillation Objective

The final Stage 3 training objective combines output-level on-policy distillation with visual- and reason-specific latent alignment:

$$\mathcal{L}_{\text{Stage3}} = \mathcal{L}_{\text{OPD}} + \lambda_v\mathcal{L}^v_{\text{latent}} + \lambda_r\mathcal{L}^r_{\text{latent}}, \tag{11}$$

where λ_v and λ_r control the strengths of visual and reasoning capability transfer. For each student on-policy rollout k of sample i, we minimize the reverse KL divergence from the student distribution p_θ to the teacher distribution p_ϕ:

$$\mathcal{L}^{i,k}_{\text{OPD}} = \frac{1}{T^k_i}\sum_{t=1}^{T^k_i} D_{\text{KL}}\left(p_\theta(\cdot \mid I_i, q_i, P, \hat{y}^k_{s,i,<t}) \,\|\, p_\phi(\cdot \mid I_i, q_i, P, \hat{y}^k_{s,i,<t})\right), \tag{12}$$

In practice, we approximate this divergence using the student's top-100 tokens and aggregate the probability mass of all remaining tokens into a single tail bucket for both distributions. We average the loss over valid response tokens within each trajectory and then over valid trajectories in the PPO mini-batch to obtain $\mathcal{L}_{\text{OPD}}$. Each loss term is averaged over its participating samples or rollouts, jointly aligning the student's on-policy predictions and visual-reason representations with the teacher.

4 Experiments

We evaluate CD-OPLD through a comprehensive comparison with distillation baselines across seven benchmarks (Section 4.2), ablations of alternative latent matching strategies (Section 4.3) and individual training components (Section 4.4), and an error-pattern analysis that examines which failures are corrected by our method (Section 4.5). The experimental setup is detailed in Section 4.1.

4.1 Experimental Setup

Model Training Our main experiments use Qwen3.5-9B as the teacher and Qwen3.5-2B as the student Qwen Team (2026). All experiments ran on 4*NVIDIA H100 80GB GPUs. In Stage 1, we perform structured cold-start SFT on a 4K subset of ViRL39K Wang et al. (2026) (3,600 training and 400 validation samples) for up to four epochs with a learning rate of 1×10^{-4}. We train LoRA adapters in the student LLM and visual merger while freezing the vision tower. In Stage 2, we train the visual and reasoning projectors on a disjoint 8K ViRL39K subset for 20 epochs with a learning

Table 1: Comparison across seven benchmarks (%). Each dataset cell reports acc@5 (top) and best@5 (bottom). The Overall AVG column reports the mean best@5 across all seven benchmarks (top) and CD-OPLD's relative improvement over the corresponding baseline (bottom). Bold denotes the best available student result.

Method	Visual Perception and Reasoning			General Visual Understanding			In-domain	Overall
	LogicVista	CV-Bench-2D	CRPE	MMStar	POPE	MME	ViRL8K	AVG best@5 / Δ_{rel}
Qwen3.5-2B (S)	25.55 64.88	64.49 89.29	46.93 67.16	53.60 82.53	87.00 97.13	74.95 96.76	37.45 73.50	81.61 ↑ 5.21%
Off-policy KD	27.16 66.67	63.73 89.57	47.72 67.85	53.20 83.00	85.96 96.96	75.28 96.84	36.75 71.50	81.77 ↑ 5.00%
OPD (w cold-start)	37.76 70.69	74.73 90.82	57.88 **72.33**	60.56 83.73	89.50 95.85	77.84 94.06	**53.75** 83.00	84.35 ↑ 1.79%
CompoDistill	30.74 70.69	70.35 **91.59**	46.36 64.57	54.09 82.73	87.74 **98.01**	72.27 **97.85**	35.70 70.75	82.31 ↑ 4.31%
LaViT	31.36 70.69	67.48 91.17	42.44 59.82	51.63 **86.13**	88.71 96.33	**78.68** 94.06	32.05 69.25	81.06 ↑ 5.92%
CD-OPLD	**38.88** **76.73**	**75.73** 91.52	**58.17** 72.28	**60.84** 84.60	**89.51** 96.02	78.31 94.86	53.35 **85.00**	**85.86** —
Teacher reference								
Qwen3.5-9B (T)	50.43 79.42	71.42 90.89	58.41 78.14	67.29 87.87	88.27 95.79	82.08 95.32	65.05 87.00	87.78 —

rate of 1×10^{-4}; only the two projectors are updated, while both model backbones remain frozen. In Stage 3, we initialize the student and projectors from their best checkpoints and perform on-policy latent distillation on the same 8K subset for two epochs with a learning rate of 1×10^{-6}. We fully update the student LLM and visual merger while keeping the teacher, projectors, and vision tower frozen. Appendix B provides the complete training configuration. Details of construction of the disjoint 4K and 8K subsets of ViRL39K can be found in appendix C.

Dataset & Evaluation Metrics We evaluate CD-OPLD on seven benchmarks spanning three categories. **Visual Perception and Reasoning** comprises LogicVista Xiao et al. (2024), CV-Bench-2D Tong et al. (2024), and CRPE Wang et al. (2024), which respectively assess visual logical reasoning, 2D spatial understanding and object counting, and inter-object relation comprehension. **General Visual Understanding** comprises MMStar Chen et al. (2024), POPE Li et al. (2023), and MME Fu et al. (2026), covering vision-dependent multimodal understanding, object hallucination, and comprehensive perceptual and cognitive abilities. We additionally evaluate in-domain multimodal reasoning on the ViRL8K test set Wang et al. (2026).

We report **acc@k** and **best@k** as our primary metrics, with $k = 5$ in the main experiments. For each test example, acc@k averages correctness over k independently sampled rollouts and then across the test set, whereas best@k assigns a score of 1 when at least one of the k rollouts is correct and 0 otherwise. For CRPE, we follow its base-question evaluation protocol: a prediction is counted as correct only when all four rotated variants of the same base question are answered correctly.

Comparison Methods For a controlled comparison, all baselines use the same 7,200/400 ViRL8K train/validation split, Qwen3.5-2B/9B student–teacher pair, two distillation epochs, and a global batch size of 128. We optimize all methods with AdamW at a learning rate of 1×10^{-6}. We compare against off-policy KD Hinton et al. (2015), OPD with the same structured cold-start initialization as CD-OPLD (**OPD w/ cold-start**) Agarwal et al. (2024), CompoDistill Kim et al. (2026), and LaViT Wu et al. (2026). Appendix F provides implementation details for all baselines.

4.2 Main Results

Table 1 shows that CD-OPLD achieves the highest acc@5 on five of the seven benchmarks, demonstrating broad gains across visual perception, relational reasoning, and general visual understanding. The exceptions are limited: LaViT is marginally better than CD-OPLD on MME acc@5 (78.68 vs. 78.31). Relative to OPD w/ cold-start, CD-OPLD increases the overall average best@5 from 84.35

Table 2: Ablation of latent matching strategies across seven benchmarks (%). Each dataset cell reports acc@5 (top) and best@5 (bottom). Bold denotes the best result for each metric.

Matching Strategies	Visual Perception and Reasoning			General Visual Understanding			In-domain
	LogicVista	CV-Bench-2D	CRPE	MMStar	POPE	MME	ViRL8K
Whole Span Matching	38.30	75.24	58.09	60.76	**89.68**	77.97	**54.30**
	76.06	90.96	72.23	83.53	95.83	94.06	84.25
Same Position Matching	38.12	74.80	56.31	59.60	89.35	78.18	52.95
	72.04	91.24	71.33	83.20	**96.08**	94.40	84.50
Semantic Block Matching (ours)	**38.88**	**75.73**	**58.17**	**60.84**	89.51	**78.31**	53.35
	76.73	**91.52**	**72.28**	**84.60**	96.02	**94.86**	**85.00**

to 85.86, a 1.79% relative improvement. The largest gain appears on LogicVista, where best@5 rises by 6.04 points (70.69 to 76.73), indicating stronger visual logical reasoning. CD-OPLD also exceeds off-policy KD, CompoDistill and LaViT in overall average best@5 by 5.00%, 4.31% and 5.92%, respectively. These results show that capability-specific latent supervision generally complements output-level OPD, improving both out-of-domain generalization and the probability of producing a correct response across multiple sampled rollouts.

4.3 Matching Strategy Ablation

We compare semantic block matching with two direct alignment strategies. **Whole Span Matching** separately mean-pools all token embeddings within the visual and reason spans of the student on-policy rollout and the teacher reference trajectory, and aligns the resulting span-level representations. **Same Position Matching** instead aligns token embeddings at identical positions within the corresponding visual and reasoning spans; for each span, alignment is restricted to the shorter of the student and teacher span lengths.

As shown in Table 2, Semantic Block Matching achieves the best acc@5 on five of the seven benchmarks and the best best@5 on six. Unlike Whole Span Matching, which compresses each span into a single representation and may dilute localized visual evidence or intermediate reasoning steps, our method preserves fine-grained semantic units and aligns only high-confidence correspondences. It also avoids the rigid one-to-one assumption of Same Position Matching, which may pair unrelated states when student and teacher trajectories differ in wording, length, or granularity. Monotonic many-to-many matching can instead align semantically equivalent blocks while skipping unmatched content, yielding more robust latent transfer across perception, relation understanding, general visual understanding, and in-domain reasoning.

The exceptions are limited: Whole Span Matching attains higher acc@5 on POPE (89.68 vs. 89.51) and ViRL8K (54.30 vs. 53.35), while Same Position Matching is marginally better on POPE best@5 (96.08 vs. 96.02). These isolated gains suggest that global or positional alignment can help on particular tasks, but the broader advantage of Semantic Block Matching supports semantic correspondence as the more reliable alignment principle.

4.4 Component Ablations

To assess the contribution of each training component, we select one representative benchmark from each evaluation category: LogicVista for Visual Perception and Reasoning, MMStar for General Visual Understanding, and ViRL8K for in-domain evaluation. We compare the full SFT+OPD+LD pipeline with three variants: SFT cold-start only, SFT cold-start with OPD, and OPD+LD. Because projector pretraining is a prerequisite for mapping student representations into the teacher latent space, we treat it as an integral part of LD rather than ab-

Table 3: Component ablations evaluated by answer accuracy (%). Green values indicate the relative improvement of the full SFT+OPD+LD pipeline over each ablated variant.

Ablation Variants	Visual Perception and Reasoning		General Visual Understanding		In-domain	
	LogicVista		MMStar		ViRL8K	
SFT Cold-start only	43.40	↑ 9.29%	63.80	↑ 0.83%	56.25	↑ 3.11%
SFT Cold-start + OPD	45.86	↑ 3.42%	62.67	↑ 2.65%	55.00	↑ 5.45%
OPD+LD	38.93	↑ 21.83%	62.20	↑ 3.42%	54.00	↑ 7.41%
SFT+OPD+LD (ours)	**47.43**	—	**64.33**	—	**58.00**	—

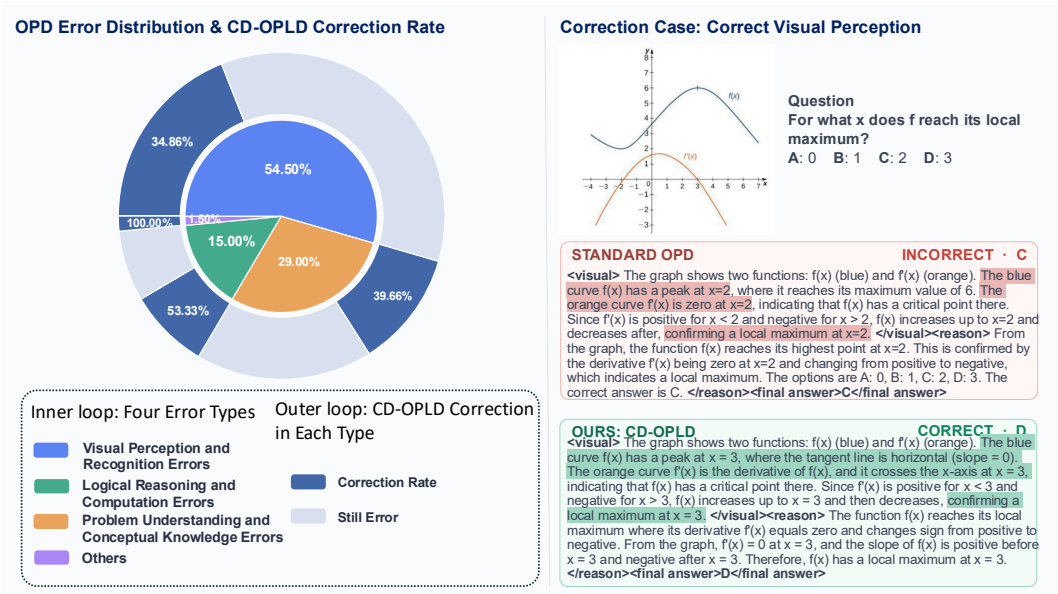

Figure 2: Distribution of OPD error types, CD-OPLD correction rates, and a representative correction case.

lating it independently. In Table 3, we report the answer accuracy from a single generated response per test example with greedy decoding for all benchmarks. Δ_{rel} measures the relative improvement of the full pipeline over each variant. Adding LD to SFT+OPD yields relative gains of 3.42%, 2.65%, and 5.45% on LogicVista, MMStar, and ViRL8K, respectively, while introducing the SFT cold start over OPD+LD improves the three benchmarks by 21.83%, 3.42%, and 7.41%. The full pipeline also consistently outperforms SFT alone, achieving gains of 9.29%, 0.83%, and 3.11%. Overall, every ablated configuration underperforms SFT+OPD+LD across all three benchmarks, demonstrating that the three components are jointly necessary for the strongest and most consistent performance.

4.5 Error Pattern Analysis

To examine which failures are addressed by capability-specific latent distillation, we randomly stratify 200 cases across the seven benchmarks for which OPD trained student is incorrect while the teacher is correct. As shown by the inner ring of Figure 2, **visual perception and recognition errors** are the dominant failure type (54.5%), followed by **problem understanding and conceptual knowledge errors** (29.0%), **logical reasoning and computation errors** (15.0%), and **other errors** (1.5%). The outer ring shows that CD-OPLD corrects 34.86%, 39.66%, 53.33%, and 100% of these four categories, respectively, yielding an overall correction rate of 40%. Appendix G reports the dataset-wise sample distribution and correction rates. The right panel presents an MMStar case in which standard OPD incorrectly locates the maximum at $x = 2$, whereas CD-OPLD correctly grounds the zero crossing of $f'(x)$ at $x = 3$ and recovers the correct answer. The results indicate that visual perception remains the primary bottleneck and that explicit visual-reason latent alignment can correct failures beyond output-level OPD.

Conclusions

We present CD-OPLD, which complements output-level OPD with explicit latent supervision that aligns the visual grounding and reasoning capabilities of student and teacher VLMs. Across seven benchmarks, CD-OPLD consistently outperforms other knowledge distillation baselines, while the ablation and error analyses further support the effectiveness of capability specific latent alignment. Nevertheless, our method requires white-box access to teacher representations and incurs additional computation for semantic matching. Its effectiveness across broader model families also remains to be investigated.

ETHICS STATEMENT

This work studies knowledge distillation for vision-language models using publicly available models and benchmarks in accordance with their respective licenses and intended research uses. It does not involve human-subject experiments or the collection of private or personally identifiable information. As with the teacher models and training data on which it depends, the distilled student may inherit social biases, factual errors, visual hallucinations, or unsafe behaviors; improved benchmark accuracy should therefore not be interpreted as eliminating these risks. The resulting models should be evaluated for fairness, robustness, privacy, and safety before deployment, particularly in high-stakes settings. We also acknowledge the computational and environmental costs of training and evaluating multiple vision-language models and report the experimental configurations to support transparent assessment and reproducibility. We are not aware of any conflicts of interest or additional ethical concerns specific to this study.

REPRODUCIBILITY STATEMENT

We provide the algorithmic formulation and training objectives of CD-OPLD in Sections 3.1–3.3, including teacher-reference selection, semantic block matching, latent alignment, and output-level OPD. Section 4.1 specifies the model pair, datasets, evaluation metrics, and shared comparison protocol. The appendix reports the complete hyperparameter and optimization configurations (Appendix B), construction of the disjoint ViRL39K training subsets (Appendix C), implementation details of all baselines (Appendix F), and the dataset-wise error-sample distribution (Appendix G). Our implementation code, training and evaluation scripts, and configuration files will be included in the supplementary materials to facilitate reproduction of the reported results.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the writing of the paper. We have not used generative AI tools for designing or providing feedback on research methodology or experiments and implementing methods. Generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims are not applicable to this work. Additionally, we used generative AI tools for summarizing or analysing existing literature, discovering research topics or identifying gaps, brainstorming, sourcing/searching for information. We have reviewed all AI-assisted work. 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*, volume 2024, pp. 21246–21263, 2024.

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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

William A Gale and Kenneth Church. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102, 1993.

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.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014. IEEE, 2025.

Jiwan Kim, Kibum Kim, Sangwoo Seo, and Chanyoung Park. Compodistill: Attention distillation for compositional reasoning in multimodal llms. In *International Conference on Learning Representations*, volume 2026, pp. 71814–71844, 2026.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

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

Yijun Liang, Yunjie Tian, Yijiang Li, Yuqi Jia, Furong Huang, Tianyi Zhou, and Di Fu. Visual contrastive self-distillation. *arXiv preprint arXiv:2607.21556*, 2026.

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.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015. URL `https://arxiv.org/abs/1412.6550`.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *Advances in Neural Information Processing Systems*, 38:30865–30891, 2026.

Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. In *European Conference on Computer Vision*, pp. 471–490. Springer, 2024.

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.

Linquan Wu, Tianxiang Jiang, Yifei Dong, Haoyu Yang, Fengji Zhang, Shichaang Meng, Ai Xuan, Linqi Song, and Jacky Keung. Lavit: Aligning latent visual thoughts for multi-modal reasoning. *arXiv preprint arXiv:2601.10129*, 2026.

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

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.

Kangning Zhang, Yixing Li, Shuai Shao, Qingyao Li, Zhengxi Lu, Zhiyuan Yao, Jianghao Lin, Wenxiang Jiao, Yuan Lu, Weiwen Liu, et al. Vad: Attributing visual evidence for target reconstruction in multimodal on-policy distillation. *arXiv preprint arXiv:2607.28590*, 2026a.

Songming Zhang, Xue Zhang, Tong Zhang, Bojie Hu, Yufeng Chen, and Jinan Xu. Kdflow: A user-friendly and efficient knowledge distillation framework for large language models. *arXiv preprint arXiv:2603.01875*, 2026b.

Shoutai Zhu, Tianyang Xu, Bin Sun, Mingyuan Xu, Yu Liu, and Qinzhen Guo. Opld: On-policy latent distillation for multimodal reasoning. *arXiv preprint arXiv:2607.28154*, 2026.

APPENDIX CONTENTS

Each entry below links to the corresponding appendix section.

Appendix A: Input Prompt and Complete Three-span Output Example 14

Appendix B: Training Details 14

Appendix C: Training Datasets Construction 16

Appendix D: Details of Teacher Offline Rollouts' Completeness Distribution 16

Appendix E: Details of Semantic Block Matching 17

Appendix F: Details of Baselines 18

Appendix G: Error Samples Distribution 19

A Input Prompt and Complete Three-span Output Example

Input prompt



Input image

```
<|im_start|>user
<|vision_start|><|image_pad|><|vision_end|>
As shown in the figure, the areas of the 6 sectors in the spinner
are all equal.  If the spinner is spun once, what is the probability
that the pointer will land on the shaded area?
Analyze the image and answer in exactly this format:
<visual>Visual understanding process.</visual>
<reason>Reason process given the question and visual
understanding.</reason>
<final answer>Final answer.</final answer>
Keep the visual and reason spans concise, and always complete all
three closing tags.<|im_end|>
```

Valid three-span rollout

```
<|im_start|>assistant
<visual>The image shows a circle divided into 6 equal sectors, alternating between black
(shaded) and white sections.  A pointer arrow indicates the start position.</visual>
<reason>Since there are 6 equal sectors and 3 are shaded, the probability of landing on a
shaded area is 3/6, which simplifies to 1/2.</reason>
<final answer>1/2</final answer>
<|im_end|>
```

Figure 3: Example input prompt and a complete, format-valid three-span output rollout.

B Training Details

For consistency with the method description, we refer to the three training stages as Stages 1–3; they correspond to Stages 0–2 in our implementation. The complete training configurations of the three stages are represented in Tables 4, 5 and 6 respectively.

Table 4: Stage 1 training configuration: structured SFT cold start.

Parameter	Value	Parameter	Value
Initial student	Qwen3.5-2B	Data split	ViRL39K_4K: 3,600 train / 400 validation
Maximum epochs	4	Optimizer	AdamW
Learning rate	1×10^{-4}	Adam betas	$(0.9, 0.95)$
Weight decay	0.01	Scheduler	Cosine decay with 20 warmup steps
Global batch size	32 samples	GPUs	4*NVIDIA H100 80GB
Micro-batch size	1 sample per GPU	Grad accumulation	8 micro-batches
Gradient clipping	1.0	Seed	42
Max sequence length	12,288	Max output length	4,096
LoRA rank / alpha	32 / 64	LoRA dropout	0.05
Trainable parameters	LoRA for student attention, linear attention, and MLP; LoRA for visual-merger fc1/fc2	Frozen parameters	Vision tower
Loss aggregation	Mean over response tokens per sample, then mean over samples in the global batch	Loss mask	Prompt and image tokens excluded
Validation frequency	Every 0.5 epoch	Early stop patience	2
ckpt policy	Permanently retain the best-accuracy checkpoint and the three most recent validation checkpoints		

Table 5: Stage 2 training configuration: capability-specific projector pretraining.

Parameter	Value	Parameter	Value
Training data	ViRL39K 8K subset	Student source	Stage 1 best ckpt
Teacher	Frozen Qwen3.5-9B	Student / teacher layer	20 / 28
Student hidden dimension	2,048	Teacher / projected dimension	4,096
Sample-group batch size	64	Temperature	0.1
Optimizer	AdamW	Learning rate	1×10^{-4}
Weight decay	0.01	Scheduler	Cosine decay with 100 warmup steps
Maximum epochs	20	Early stop patience	3 epochs
Gradient clipping	1.0	Initialization	Xavier uniform
Seed	42	Trainable parameters	Visual and reason projectors only
Frozen parameters	Student and teacher backbones	Loss aggregation	Mean over correct-rollout anchors within each modality, then mean over valid sample groups in the batch

Table 6: Stage 3 training configuration: on-policy output and latent distillation.

Parameter	Value	Parameter	Value
Training data	ViRL39K 8K subset	Student initialization	Stage 1 best ckpt
Teacher	Frozen Qwen3.5-9B	Projectors	Frozen Stage 2 best ckpt
Trainable parameters	Full student LLM and visual merger	Frozen parameters	Vision tower, teacher, and projectors
Train batch size	128 samples	PPO mini-batch size	128 trajectories
Rollouts per sample	1	PPO epochs per batch	1
Optimizer	AdamW	Learning rate	1×10^{-6}
Adam betas	$(0.9, 0.999)$	Weight decay	0.01
Scheduler / warmup	Constant / 5 steps	Gradient clipping	1.0
Training epochs	2	Total optimizer steps	112
Drop last	True	Seed	42
Rollout sampling	Temperature 1.0; top-p 1.0; no top-k limit; repetition penalty 1.0	Length limits	Prompt 8,192; response 4,096; total 12,288 tokens
OPD approximation	Student top-100 tokens plus one aggregated tail bucket	OPD aggregation	Mean over response tokens per trajectory, then mean over valid PPO mini-batch trajectories
Student / teacher latent layer	20 / 28 for both capabilities	Latent-loss weights	$\lambda_v = 0.10$, $\lambda_r = 0.10$
Matching threshold	0.92 for visual and reasoning spans	Maximum grouped chunks	3 student chunks; 2 teacher chunks
Matching-score weights	Hidden 0.50; visual anchor 0.25; equation 0.25	Alignment penalties	Student gap 0.05; teacher gap 0.05; group size 0.05
Initial validation	Step 0	Validation frequency	Every 14 optimizer steps
Ckpt save freq	Every 28 optimizer steps	Early stopping	Disabled
Ckpt policy	Permanently retain best checkpoint and three latest periodic checkpoints	GPUs / FSDP ranks	4 / 4

C Training Datasets Construction

We construct two mutually disjoint subsets from ViRL39K Wang et al. (2026): a 4K subset for the structured SFT cold start and an 8K subset for projector pretraining and on-policy latent distillation. Both subsets are sampled by jointly stratifying the eligible ViRL39K examples by task category and pass-rate-based difficulty. The 4K subset is further split into 3,600 training and 400 validation examples. The 8k subset is further split into 7,200 training, 400 validation and 400 test examples.

Pass-rate-based difficulty tiers. The difficulty labels are not manually annotated. Instead, we derive them from the two model pass-rate fields provided by ViRL39K. Let p_{32} = `PassRate_32BTrained` and p_7 = `PassRate_7BBase`. We assign each eligible example to one of three difficulty tiers as

$$\text{Difficulty}(x) = \begin{cases} \text{Easy}, & p_{32} = 1 \text{ and } p_7 \geq 0.75, \\ \text{Medium}, & p_{32} = 1 \text{ and } 0 < p_7 < 0.75, \\ \text{Hard}, & 0 < p_{32} < 1 \text{ or } (p_{32} = 1 \text{ and } p_7 = 0). \end{cases} \tag{13}$$

Thus, Easy examples are solved consistently by the trained 32B model and at least 75% of the time by the base 7B model. Medium examples are always solved by the trained 32B model but only partially solved by the base 7B model. Hard examples either cannot be solved at all by the base 7B model or cannot be solved consistently even by the trained 32B model. Examples with $p_{32} = 0$, as well as examples not covered by Equation 13, are excluded from the sampling pool.

Category-aware allocation. ViRL39K contains eight task categories: (1) grade-school geometric reasoning, (2) grade-school non-geometric mathematics, (3) grade-school science, (4) broader STEM topics, (5) commonsense, (6) social science, (7) spatial reasoning, and (8) tables, diagrams, and charts. To prevent large categories from dominating while retaining the diversity of the original corpus, half of each subset is allocated uniformly across categories and the other half proportionally to the category frequencies in the eligible pool. For a subset of size n, the quota for category c is

$$N_c(n) = \frac{1}{2}\frac{n}{8} + \frac{1}{2}\left(n\frac{M_c}{M}\right), \tag{14}$$

where M_c is the number of eligible ViRL39K examples in category c and $M = \sum_c M_c$. We use $n = 8{,}000$ for the ViRL8K subset and $n = 4{,}000$ for the SFT subset.

Joint category–difficulty stratification. Within the category quotas, we further stratify examples into the Easy, Medium, and Hard tiers, producing $8 \times 3 = 24$ category–difficulty strata. For ViRL8K, the global difficulty totals are constrained to

$$(N_{\text{Easy}}, N_{\text{Medium}}, N_{\text{Hard}}) = (0.15, 0.45, 0.40) \times 8{,}000 = (1{,}200, 3{,}600, 3{,}200). \tag{15}$$

The same proportions are used for the 4K subset, corresponding to 600 Easy, 1,800 Medium, and 1,600 Hard examples. We sample the two subsets without overlap, so the structured SFT cold-start data are disjoint from all examples used in the subsequent projector-pretraining and on-policy-distillation stages.

D Details of Teacher Offline Rollouts' Completeness Distribution

For every sample in the ViRL8K train/validation split, the teacher generates five offline rollouts. A rollout passes the completeness filter only if it contains exactly the three canonical spans, `<visual>`, `<reason>`, and `<final answer>`; all three spans must be non-empty and correctly closed, and none of their contents may reach the corresponding tokens length limit, which would indicate possible truncation. Table 7 reports the distribution of complete teacher offline rollouts across the ViRL8K train and validation splits.

Table 7: Distribution of complete teacher offline rollouts across the ViRL8K train and validation splits. Five rollouts are generated per sample.

Complete rollouts per sample	Train	Val	Overall
0	575	56	631
1	485	35	520
2	555	42	597
3	662	39	701
4	1,189	68	1,257
5	3,734	160	3,894
Total samples	**7,200**	**400**	**7,600**

Among the 7,600 samples, 3,894 (51.24%) retain all five complete rollouts, whereas 6,969 (91.70%) retain at least one. The remaining 631 samples (8.30%) have no rollout that passes the filter. Overall, the filter retains 28,315 of the 38,000 generated rollouts (74.51%).

Table 8 further partitions the samples according to whether their complete rollouts contain a correct answer. The 5,753 samples with at least one complete and correct rollout provide student anchors, their paired teacher positives, and teacher negatives. The 1,216 samples that contain complete rollouts but no correct rollout cannot provide anchors; their complete teacher rollouts are used only as cross-sample negatives. The 631 samples without any complete rollout are excluded from projector pretraining.

Table 8: Sample distribution used for capability-specific projector pretraining.

Sample type	Train	Val	Overall
At least one complete, correct rollout	5,467	286	5,753
Complete rollout(s), but none correct	1,158	58	1,216
No complete rollout	575	56	631
Total	**7,200**	**400**	**7,600**

E Details of Semantic Block Matching

Span parsing and recovery. The selected teacher reference is parsed strictly and must contain the complete ordered sequence `<visual>...</visual>`, `<reason>...</reason>`, and `<final answer>...</final answer>`. Student on-policy responses are parsed permissively: complete visual and reasoning spans are extracted when available, content associated with incomplete tags is recovered whenever its span type can be inferred, and ambiguous text is retained as *untyped*. The final-answer span is excluded from latent matching on both sides.

Semantic chunking. We segment each recovered visual or reasoning span at sentence-ending punctuation, line breaks, mathematical-expression boundaries, and list structure. The rules avoid splitting decimal numbers, abbreviations such as *e.g.* and *i.e.*, option markers such as "A.", multiline formulas, isolated connectives, and formula continuations. Each chunk retains its exact character interval, which is subsequently mapped to the corresponding response-token range.

Many-to-many candidates and teacher-space representations. Visual and reasoning spans are processed separately. A candidate block may contain one to three consecutive student chunks and one or two consecutive teacher chunks, supporting 1–1, 2–1, 3–1, 1–2, 2–2, and 3–2 correspondences. We teacher-force both trajectories through the frozen teacher and mean-pool its layer-28 token hidden states over each candidate block. Thus, all semantic similarities are computed in the same representation space despite differences between the student and teacher backbones.

Matching features and conflict filtering. The semantic feature is the nonnegative cosine similarity between the pooled teacher-space representations. For the visual-anchor feature, we extract geometric entities, diagram labels, and spatial relations, such as `AB`, $\angle ABC$, triangle, radius, paral-

lel, left, and above. For the equation feature, we compare normalized equation skeletons containing variables, number placeholders, operators, and mathematical concepts, such as `NUM`, x, $=$, $+$, $-$, radius, area, and `\frac`. Let $\mathcal{A}(b)$ and $\mathcal{E}(b)$ denote the resulting anchor and equation-feature sets. Their similarities are

$$s_a(b) = \frac{|\mathcal{A}(b_s) \cap \mathcal{A}(b_t)|}{|\mathcal{A}(b_s) \cup \mathcal{A}(b_t)|}, \qquad s_e(b) = \frac{|\mathcal{E}(b_s) \cap \mathcal{E}(b_t)|}{|\mathcal{E}(b_s) \cup \mathcal{E}(b_t)|}. \tag{16}$$

Each feature is included only when its required inputs are available and its set union is nonempty; the weights of the remaining features are renormalized. We use weights 0.50, 0.25, and 0.25 for the semantic, visual-anchor, and equation features, respectively. Before scoring, we reject candidates with explicit semantic conflicts, including different answer choices, incompatible numerical assignments to the same variable, or clearly inconsistent final numerical conclusions.

Monotonic alignment and malformed responses. Only candidates with $s_b \geq 0.92$ may enter the matching process. We set the student-gap, teacher-gap, and group-size penalties to 0.05 each. The dynamic program may match a candidate block, skip one student chunk, or skip one teacher chunk, but the selected blocks must preserve trajectory order and cannot overlap. Untyped student chunks are matched against the combined teacher visual and reasoning sequences, but no candidate may cross the teacher's visual–reason boundary. An accepted untyped match inherits the teacher-side span type and contributes to the corresponding visual or reasoning latent loss.

F Details of Baselines

Off-policy KD. Classical knowledge distillation transfers a teacher's predictive knowledge to a compact student by matching the teacher's softened output distribution on a fixed training corpus Hinton et al. (2015). In our implementation, we realize this off-policy baseline using KDFlow Zhang et al. (2026b): the student and teacher are teacher-forced on the same ground-truth answer conditioned on each ViRL8K image–question pair, and the KD loss is applied only to the ground-truth answer tokens and EOS token, excluding prompt and image tokens, without any on-policy rollout during training.

Pure OPD. On-policy distillation mitigates the train-inference distribution mismatch of autoregressive models by training the student on its own generated sequences and obtaining token-level teacher feedback on the states it actually visits Agarwal et al. (2024). In our implementation, Pure OPD samples one student on-policy rollout for each ViRL8K image-question prompt and minimizes the student to teacher reverse KL on the student-generated prefixes, using the student's top-100 tokens plus an aggregated tail bucket and averaging first over eligible response tokens and then over the 128 trajectories; tokens inside the `visual`, `reason`, and `final answer` spans participate, delimiter, prompt, and image tokens are masked, malformed responses retain all non-delimiter response tokens.

CompoDistill. CompoDistill identifies visual-attention misalignment between student and teacher as a key obstacle to transferring visual perception and explicitly aligns their visual attention to improve compositional multimodal reasoning Kim et al. (2026). In our implementation, CompoDistill is trained with teacher forcing on a constructed response containing the ground-truth answer. During the first epoch (56 steps), its objective assigns equal weights to full-response CE, teacher-to-student forward KL approximated using the teacher's top-100 tokens and an aggregated tail bucket, and a visual-attention alignment loss. Specifically, the visual-token attention maps from student layers $\{3, 7, 11, 15, 19, 23\}$ are aligned by cosine loss with the mean attention maps of the corresponding teacher-layer triplets $\{3, 7, 11\}$, $\{7, 11, 15\}$, $\{11, 15, 19\}$, $\{15, 19, 23\}$, $\{19, 23, 27\}$, and $\{23, 27, 31\}$, respectively, with attention-sink filtering applied. Training then switches to CE-only SFT for the second epoch (56 steps).

LaViT. LaViT addresses the perception gap between teacher and student by aligning latent visual representations with the teacher's visual semantics and attention trajectories, while using curriculum sensory gating to discourage bypassing the latent representations Wu et al. (2026). In our implementation, LaViT is adapted to a Qwen3.5-2B student with a frozen Qwen3.5-9B teacher, inserting four learned latent tokens before the supervised response. Across both training epochs (112 steps), the

objective combines CE with semantic reconstruction and trajectory alignment losses, weighted by 1, 0.3, and 0.3, respectively. Specifically, the final latent hidden state is projected to reconstruct the teacher's mean final-layer visual representation through cosine loss and predict a patch distribution aligned through forward KL with the teacher's gaze, aggregated across its eight full-attention layers and sparsified to the top eight patches. The sensory gate follows a cosine schedule from 10^{-6} to 1 over the first 45 steps. To accommodate Qwen3.5's hybrid architecture, full-attention layers apply a text-to-image attention bias, while linear-attention layers blend standard outputs with outputs computed after excluding image tokens, preserving unrestricted visual access for latent queries.

G Error Samples Distribution

Table 9 reports the dataset-wise distribution of the 200 error cases analyzed in Section 4.5. We first form an eligible pool containing examples for which the OPD-trained student answers incorrectly while the teacher answers correctly, and then sample from each benchmark in proportion to its share of this pool using seed 46. For CRPE, evaluation is performed at the base-question level: a model is considered correct only if it answers all four rotated variants correctly. CD-OPLD corrects 80 of the 200 sampled errors overall (40.00%), with dataset-level correction rates ranging from 26.32% on CRPE to 71.43% on ViRL8K.

Table 9: Dataset-wise distribution and CD-OPLD correction results for the 200 stratified error samples selected with seed 46.

Dataset	Number of Samples	Ours Correct Number	Ours Correction Rate
LogicVista	15	5	33.33%
CV-Bench-2D	18	9	50.00%
CRPE	38	10	26.32%
MMStar	42	17	40.48%
POPE	31	13	41.94%
MME	42	16	38.10%
ViRL8K	14	10	71.43%
Total	200	80	40.00%