SELF-SUPERVISED VISUAL ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual on-policy distillation relies heavily on an informative teacher–student asymmetry, through either a larger, stronger teacher or privileged supervision, such as reference answers or ground-truth regions of interest. This raises a fundamental question: *where can informative asymmetry come from when nothing privileged is available?* We answer this by inverting where the asymmetry comes from. Rather than adding privileged information to the teacher, we subtract information from the student. This asymmetry creates the same effective learning signal *for free* as a teacher with access to information unavailable to the student, without ground-truth annotations, rewards, or a separate stronger teacher model. Building on this principle, we introduce **Self-Supervised Visual On-Policy Distillation** (S^2VOPD), a simple yet effective method that constructs on-policy learning signals from asymmetric augmented views. S^2VOPD distills the teacher's distribution conditioned on the original image on-policy into the student distribution conditioned on a strongly augmented view of the same image. We systematically explore a broad design space of visual augmentations and uncover that (1) *asymmetry*: all four augmentation families improve performance, while symmetric self-distillation degrades it; (2) *moderate strength*: performance peaks at a moderate strength; and (3) *composition*: multiple strong augmentations composed outperforms the best single augmentation. Across six fine-grained perception benchmarks, S^2VOPD improves Qwen3.5-4B from 70.7% to 77.8% and Qwen3.5-9B from 73.4% to 78.8%—above all open-source models compared, up to Qwen3.5 at 397B, and surpassing GPT-5.4. While holding training data the same, it also outperforms all methods trained with privileged information such as ground-truth regions or answers and self-rewarding method without any annotation.

1 INTRODUCTION

On-policy distillation (OPD) has been shown to be effective in post-training of both LLMs (Agarwal et al., 2024; Zhao et al., 2026a) and VLMs (Bousselham et al., 2026; Liu et al., 2026). It improves a student policy by providing dense teacher supervision along the student's own generated trajectories (Agarwal et al., 2024; Zhao et al., 2026a; Sang et al., 2026). Central to this paradigm is an informative *teacher–student asymmetry*: the teacher must possess information or capability that enables it to provide a better predictive target than the student itself. Conventional OPD obtains this asymmetry from a larger or stronger teacher. On-policy self-distillation (OPSD) removes the need for a separate teacher model, but instead conditions the teacher on privileged information, such as reference solutions (Zhao et al., 2026a; Penaloza et al., 2026), environmental feedback (Hübotter et al., 2026), or ground-truth regions of interest (Yuan et al., 2026; Liu et al., 2026; Yoon et al., 2026). In either case, the learning signal ultimately relies on a privileged advantage external to the model, which becomes increasingly difficult to obtain as model capabilities outpace the supervision humans can reliably provide.

This raises a fundamental question:

where can informative asymmetry come from when nothing privileged is available?

We show that it does not necessarily come from giving the teacher more information at all. Instead, we invert the direction of the asymmetry: *rather than adding information to the teacher, we withhold information from the student*. The teacher observes the original input, while the student operates on a degraded view of the same input. Their difference in available visual evidence induces a

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.

predictive discrepancy and thereby creates the asymmetry required for distillation—without ground-truth annotations, rewards, privileged context, or a separately trained teacher. This idea is similar to self-supervised learning, where supervision is constructed from multiple views of the same unlabeled input (Tian et al., 2020; Chen & He, 2021; Tian et al., 2025; Zaher et al., 2026; Sharma et al., 2026). Contrastive methods such as SimCLR (Chen et al., 2020) learn by aligning representations of differently augmented views, while teacher–student methods such as BYOL (Grill et al., 2020) and DINO (Caron et al., 2021) use one view to provide learning targets for another augmented view. They all share the principle that augmentation defines invariances that allow models to learn without external labels. Our method follows the same principle, but uses augmentation for a different purpose. Rather than defining a representation-level objective across augmented views, we apply augmentation to invert where the asymmetry comes from, thereby creating an informative asymmetry *for free*.

Building on this idea, we introduce **Self-Supervised Visual On-Policy Distillation** (S^2VOPD). Given an image x and question q, an exponentially moving average teacher observes the original image, while the student generates on-policy trajectories from an augmented view $\tilde{x} = T(x)$. S^2VOPD then conditions the teacher distribution on the original image and distills into the student trajectories generated from an augmented view $\tilde{x} = T(x)$. This transformation T induces a predictive gap between teacher and student for free, without labels, regions of interest, or external rewards; the student-side transformation is the sole source of the teacher's informational advantage.

Once asymmetry is constructed this way, a more important question emerges: *how to construct asymmetry through augmentation?* We answer this question empirically, through what is to our knowledge the first controlled study of the student-view augmentation space for on-policy distillation. The analysis first establishes three properties of an informative single augmentation: (1) *asymmetry is what matters*: every augmentation family improves over the base model, whereas symmetric self-distillation degrades it; (2) *the perturbation must be moderate*: accuracy increases with the induced teacher–student gap up to a point and declines beyond it; and (3) *the gap must remain task-consistent*: aggressive cropping yields the largest discrepancy in our study yet the weakest supervision, because it removes the evidence the question depends on. We then turn from individual operators to their combination, and establish three further properties: (4) *composition surpasses the best individual operator*: a composition of three complementary augmentations improves over the strongest single one at both model scales; (5) *the composition must be exclusive*: sampling one operator per student view outperforms applying two strong operators to the same view, which destroys recoverable visual evidence; and (6) *weak operators are the exception*: one may be stacked onto a strong augmentation with improved performance.

These findings identify the most effective instantiation of S^2VOPD: an exclusive composition of three augmentations with probability 0.9, one of which is sampled per student view, otherwise identity. Across six fine-grained perception benchmarks, S^2VOPD improves Qwen3.5-4B from 70.7% to 77.8% in average accuracy, a gain of 7.1%. This places a 4B model above Qwen3-VL-Instruct-235B (75.8%) and GPT-5.4 (72.8%), and above the much larger Qwen3.5-397B (77.4%). Moreover, S^2VOPD consistently improves both fine-grained perception and math reasoning at both scales, by $+7.1\%$ and $+3.3\%$ at 4B and $+5.5\%$ and $+4.0\%$ at 9B, respectively, surpassing methods with privileged-supervision and self-rewarding RL methods by a large margin.

2 Related Work

Self- and semi-supervised learning. Self-supervised representation learning derives supervision from multiple augmented views of unlabeled images. Contrastive methods such as SimCLR (Chen et al., 2020) and MoCo (He et al., 2020) align different views of the same image, while negative-free methods such as DINO (Caron et al., 2021) match a student to a stop-gradient or momentum teacher across augmentations. Similar asymmetry appears in semi-supervised consistency learning, where weakly augmented predictions supervise strongly augmented inputs (Laine & Aila, 2017; Tarvainen & Valpola, 2017; Xie et al., 2020a; Sohn et al., 2020; Xie et al., 2020b; Li et al., 2023).

Visual augmentation in vision-language models. NoisyRollout Liu et al. (2025) perturbs images to diversify rollouts, while VPPO Huang et al. (2026) and PRPO Li et al. (2026) compare policy distributions across clean and perturbed views to reweight policy gradients. These methods still rely on external rewards. In contrast, we augment only the student view and use the induced asymmetry as the training signal without any GT rewards or annotations.

Knowledge and on-policy distillation. On-policy distillation (OPD) Agarwal et al. (2024) trains the student on its own trajectories while receiving dense token-level supervision from a teacher, reducing the train–inference mismatch of offline distillation. On-policy self-distillation (OPSD) extends this idea further using the model itself conditioned on ground-truth solution as the teacher. Existing methods further utilize verified reasoning traces Zhao et al. (2026a), additional context Ye et al. (2026), environmental feedback Hübotter et al. (2026), or privileged visual information Yuan et al. (2026); Liu et al. (2026); Yoon et al. (2026); Sun et al. (2026) as previlaged information.

We refer to Appendix B for full version of related work

3 METHOD

3.1 SELF-SUPERVISED VISUAL ON-POLICY DISTILLATION

Given an image–question pair $(x, q) \sim \mathcal{D}$, let π_θ denote the policy to be optimized and π_ϕ its exponential-moving-average (EMA) as teacher: $\phi \leftarrow (1-\eta)\,\phi + \eta\,\theta$. Unlike conventional on-policy distillation, π_ϕ is neither a separately trained stronger model nor conditioned on privileged supervision. Instead, we construct the teacher–student asymmetry directly from their visual inputs.

Asymmetric on-policy views. For each $(x, q) \sim \mathcal{D}$, we sample a stochastic transformation $T \sim \mathcal{T}$ that produces the student view $\tilde{x} = T(x)$, while the teacher retains the clean view x. The student samples n rollouts $y^{(1)}, \ldots, y^{(n)} \sim \pi_\theta(\cdot \mid \tilde{x}, q)$ conditioned on its corrupted view $\tilde{x}$. This gives us a teacher distribution conditioned on the clean view x: $p_t^\tau = \pi_\phi(\cdot \mid x, q, y_{<t})$ and the student distribution conditioned on the augmented view $\tilde{x}$: $p_t^s = \pi_\theta(\cdot \mid \tilde{x}, q, y_{<t})$.

On-policy distillation. With original clean view x for teacher model and the augmented view $\tilde{x}$ for the student, S^{2}VOPD then conditions the teacher's next-token distribution on the clean view x, and distills into the student's own trajectories generated on the degraded view $\tilde{x}$. Let $D(\cdot \,\|\, \cdot)$ denote a divergence between two distributions over the vocabulary. The training objective minimizes the expected per-token divergence between the teacher and the student distributions, evaluated along trajectories sampled from the student:

$$\mathcal{L}(\theta) \;=\; \mathbb{E}_{(x,q)\sim\mathcal{D}}\, \mathbb{E}_{y\sim\pi_\theta(\cdot\mid\tilde{x},q)} \left[\frac{1}{|y|} \sum_{t=1}^{|y|} D\big(\pi_\phi(\cdot \mid x,\, q,\, y_{<t}) \,\big\|\, \pi_\theta(\cdot \mid \tilde{x},\, q,\, y_{<t})\big) \right]. \tag{1}$$

The inner expectation renders the objective on-policy: the prefixes $y_{<t}$ at which the two distributions are compared are drawn from the student's own policy, so that supervision is provided precisely at the states the student visits at inference time.

We instantiate D as the generalized Jensen–Shannon divergence D_{JS}^{α}, which interpolates between the forward and reverse Kullback–Leibler divergence D_{KL} through a mixture weight α and remains bounded when the two distributions have limited overlap. With $\alpha = 0.5$ and the mixture distribution $m_t = \alpha\,\pi_\phi(\cdot \mid x, q, y_{<t}) + (1-\alpha)\,\pi_\theta(\cdot \mid \tilde{x}, q, y_{<t})$,

$$\begin{aligned} D_{\mathrm{JS}}^{\alpha}\big(\pi_\phi(\cdot \mid x,\, q,\, y_{<t}) \,\big\|\, \pi_\theta(\cdot \mid \tilde{x},\, q,\, y_{<t})\big) \;&=\; \alpha\, D_{\mathrm{KL}}\big(\pi_\phi(\cdot \mid x,\, q,\, y_{<t}) \,\big\|\, m_t\big) \\ &\quad + (1-\alpha)\, D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid \tilde{x},\, q,\, y_{<t}) \,\big\|\, m_t\big). \end{aligned} \tag{2}$$

Both distributions are restricted to the teacher's top-k tokens at position t and renormalized over this support, which stabilizes the objective against the long tail of the vocabulary.

Crucially, the transformation T constructs an input asymmetry that induces a predictive discrepancy between the student distribution p_t^s and the teacher distribution p_t^τ. Through T, the student is trained on a deliberately degraded view of the same image, while the teacher evaluates the identical on-policy prefix under the original view. Unlike conventional OPD, which obtains this asymmetry from a stronger teacher or privileged supervision, S^{2}VOPD creates it by withholding visual information from the student. This yields an analogous on-policy distillation signal for free without external labels, rewards, privileged annotations, or a stronger teacher. The effectiveness of this self-constructed supervision therefore depends on how T shapes the teacher–student predictive gap.

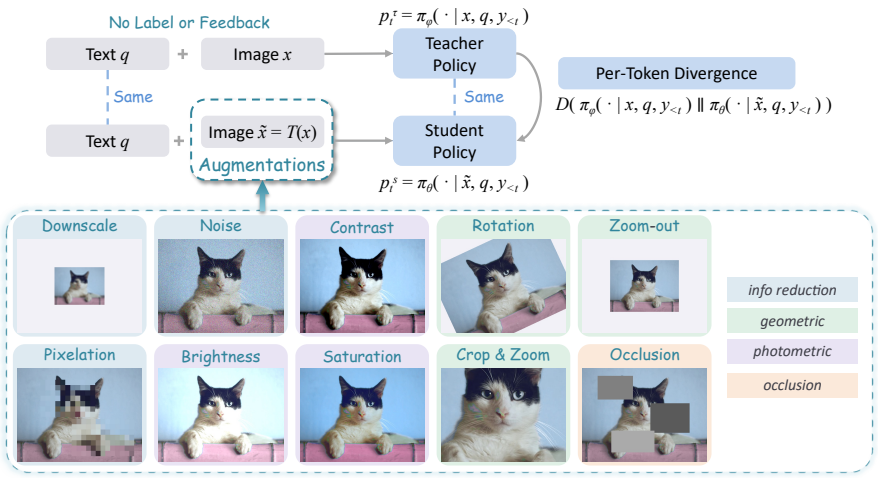

Figure 1: Overview of S^{2}VOPD. The student generates rollouts from a corrupted view of the image; an EMA teacher scores the same generated prefixes under the clean view; a top-k generalized JSD transfers the teacher's better-informed token distributions to the student. The augmentation applied to the student view is the sole source of supervision and the object of our study.

3.2 Constructing Asymmetry for Free

Augmentation space. The transformation T is specified by three components: a set of primitive operators together with their probabilities and strength ranges, a composition policy for combining selected operators, and a global probability controlling whether augmentation is applied at all. We organize these operators by how they alter the visual information available to the student, yielding four families (see Table 5 in Appendix A for details):

- **Information reduction.** Operators that reduce usable visual information while preserving spatial layout. Subtractive variants remove signal content, including downscaling, Gaussian blur, pixelation, and spectral band-stop filtering; additive variants reduce signal-to-noise ratio through Gaussian noise. We also include visual-token dropping, which reduces the number of encoder tokens while preserving the resolution and positions of those retained.
- **Geometric.** Operators that modify spatial organization, including rotation, translation, cropping, and zoom-out with padding. Unlike the other families, these transformations may alter the answer to spatially grounded questions by changing object locations relative to the image frame.
- **Photometric.** Operators that modify appearance while preserving geometry and content, including brightness, contrast, saturation, hue, gamma, sharpness, and histogram equalization. These transformations leave object boundaries and spatial relations unchanged.
- **Occlusion.** Operators that remove localized image regions while leaving the remainder unchanged, including random erasing, grid masking, and filled crops. This family is distinguished from information reduction by spatial locality. We additionally vary the fill content (e.g., constant or noise) and occlusion granularity (few large versus many small regions).

Formally, let $\mathcal{A} = \{A_1, \ldots, A_M\}$ denote the collection of candidate operators, where operator A_j is associated with an application probability ρ_j and a strength $\lambda_j \sim P_j$. We first sample a global augmentation indicator

$$z \sim \text{Bernoulli}(p), \tag{3}$$

where p controls the fraction of student views that are augmented. When $z = 0$, the student receives the original view, i.e., $T(x) = x$. When $z = 1$, we independently sample

$$b_j \sim \text{Bernoulli}(\rho_j), \qquad j = 1, \ldots, M, \tag{4}$$

and compose the selected operators in a fixed order:

$$T(x) = \begin{cases} x, & z = 0, \\ \left(A_M^{b_M}(\cdot\,; \lambda_M) \circ \cdots \circ A_1^{b_1}(\cdot\,; \lambda_1)\right)(x), & z = 1, \end{cases} \tag{5}$$

where A_j^0 denotes the identity and A_j^1 applies A_j. This formulation jointly controls the augmentation family, the strength, the per-operator probability, the composition, and the probability $1 - p$ of retaining an unaugmented view, and thereby allows us to construct teacher–student asymmetries of different forms and magnitudes in a systematic way. Throughout, augmentation is applied to the student view only; the teacher always observes the original image.

4 EXPERIMENTS

4.1 SETUP

Model and training. We adopt Qwen3.5-4B and Qwen3.5-9B (Team, 2026) as base models for all experiments. We utilize. Every run uses an identical configuration, unless specified otherwise: batch size 96 prompts, n=8 rollouts per prompt, learning rate 2×10^{-6} with 10 warmup steps, 65 optimizer steps (one epoch of the 6K samples). We apply EMA with η=0.05 on the teacher and a maximum prompt length of 8,192 and response length of 1,024 tokens. For a fair comparison, we adopt Vision-OPD-6K (Yuan et al., 2026) as the training set, whose ground-truth region annotations are only provided to Vision-OPD (Yuan et al., 2026) for training and are never used by S^{2}VOPD.

Evaluation. We evaluate on six fine-grained perception benchmarks: V*Bench (Wu & Xie, 2024), ZoomBench (Wei et al., 2026), HR-Bench 4K and 8K (Wang et al., 2025), MME-RealWorld and its Chinese subset (Zhang et al., 2025). To evaluate models' generalization to out-of-distribution domains, we further adopt three math reasoning benchmarks: MathVista (Lu et al., 2024), MathVerse (Zhang et al., 2024), and MathVision (Wang et al., 2024). For inference on the six perception benchmarks, we use greedy decoding with a maximum of 4,096 tokens; for math reasoning benchmarks, we adopt a 24,576-token budget with temperature T=0.3, top-p 0.95, top-k 20, and presence penalty 1.5. Generations are evaluated by first extracting the final answer and applying near-exact matching; only cases unresolved by these rules are adjudicated by an LLM judge (i.e., Qwen2.5-72B-Instruct).

Baselines. We compare S^{2}VOPD with various models and baselines: (i) the base model; (ii) symmetric self-distillation (w/o Aug. in Table 3) where on-policy self-distillation is performed without privileged information and augmentation; (iii) three methods with different privileged information, i.e., GT region of interest annotations (ZwZ (Wei et al., 2026), Vision-OPD (Yuan et al., 2026)), and OPSD with GT answer as privileged information for teacher, trained OPSD with Qwen3.5-4B on the Vision-OPD-6k dataset; and (iv) three self-rewarding RL methods, i.e., TTRL (Zuo et al., 2025), Intuitor (Zhao et al., 2026b), and RENT (Prabhudesai et al., 2026), all trained for 65 steps on Vision-OPD-6k dataset, following a similar configurations.

4.2 SEARCHING IN THE AUGMENTATION SPACE

Which augmentations are useful? We next ask which augmentations create an informative teacher–student asymmetry that can serve as an effective predictive learning signal. Each configuration below is a complete training run that differs only in T; the training data, optimizer, step budget, and evaluation protocol are held fixed. Due to limited compute, these analysis runs are evaluated with greedy decoding at a 2,048-token budget; their absolute numbers are therefore internally comparable but sit slightly below the 4,096-token rows of Table 1.

All augmentation families are effective. Under this fixed analysis protocol, the base model achieves 70.58%. Symmetric self-distillation with T equal to the identity reaches only 65.21%, slightly degrading the initial model. In contrast, applying each augmentation family independently yields 75.65% for information reduction, 74.40% for photometric transformations, 74.30% for geometric transformations, and 72.44% for occlusion (Figure 2(a)). Thus, every family substantially improves over both the base model and unaugmented self-distillation.

Strength matters: neither too weak nor too strong. Across all augmentation families, performance first improves and then declines as the strength increases. We show this in Figure 2(b) by plotting the averaged performance against the teacher-student gap induced by the augmentation. We measure teacher-student gap as the token-level JS Divergence between the teacher and the student distributions, averaged over the first ten steps of training, which aligns well with the strength parameters of each augmentation (e.g., heavy blur induces a larger teacher–student gap than light blur). Resolution reduction peaks at 75.65% for a scale range of 0.3–0.6, compared with 75.25% and 75.05% at weaker

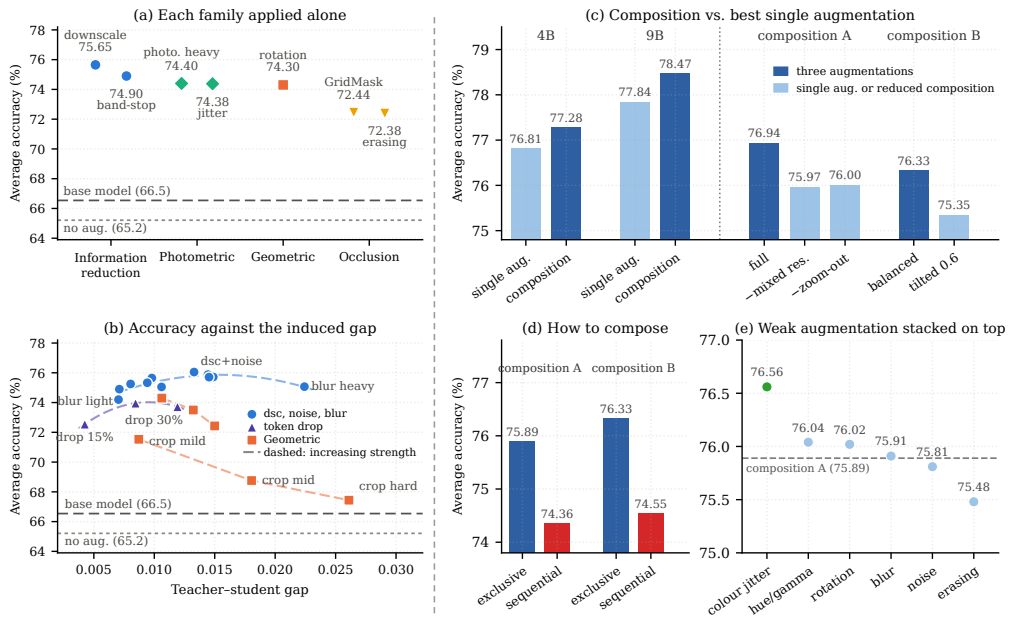

Figure 2: Analysis of the augmentation space. **(a)–(b)** Single augmentations: **(a)** performance of each augmentation family applied alone; **(b)** accuracy against the teacher–student predictive gap induced by the augmentation, measured as the token-level JS divergence between the teacher and the student distributions, averaged over the first ten steps of training. **(c)–(e)** Composing augmentations: **(c)** the three-augmentation composition (B at 4B, A at 9B) against the best single augmentation (mixed resolution) (left), and ablations at 4B (right): removing augmentations from composition A or tilting the augmentation probability weights of composition B from $0.3/0.3/0.3$ to $0.6/0.15/0.15$; **(d)** exclusive versus sequential composition of the same augmentations, for compositions A and B; **(e)** a weak augmentation stacked onto every branch of composition A, where the dashed line is composition A alone, trained in the same batch.

and stronger strength. Blur follows the same trend, rising from 74.20% to 75.71% before falling back to 75.07%, as does visual-token dropping, which moves from 72.55% up to 73.96% and back down to 73.72%. More generally, plotting accuracy against the induced teacher–student predictive gap reveals a consistent pattern: performance increases with the gap up to a JS divergence of roughly 0.014, then declines beyond it. Thus, useful asymmetry requires sufficient, but not excessive, perturbation.

A larger gap is not a better gap: the augmentation must preserve question-relevant information. Cropping exposes this limitation. Unlike information-reduction augmentations, its performance decreases monotonically with strength, falling from 71.53% to 68.76% and then 67.44% across mild, moderate, and strong crops (Figure 2(b)). The strongest crop produces the largest predictive gap in our study yet achieves only 67.44%, barely above unaugmented self-distillation. Even the mildest setting, whose induced gap coincides with that of the best resolution and blur configurations, trails them by more than 3%. The degradation is especially pronounced on benchmarks whose answers depend on localized or global visual evidence: moving from moderate to strong cropping reduces V*Bench by 2.1%, HR-Bench 8K by 2.4%, and MME-RealWorld by 2.1%. Cropping may remove the evidence required to answer the question altogether. The resulting discrepancy is therefore large partly because the student's input has become unanswerable, and such a gap provides little useful learning signal.

Overall, useful asymmetry is governed by both magnitude and semantics. The augmentation must create a sufficient information gap to induce learning, while the structure of that gap determines which perceptual capabilities are emphasized. Downscaling with Gaussian noise offers the most robust default across benchmarks.

How to compose augmentations? Having identified which augmentation families create useful asymmetry, we ask whether several effective augmentations can be combined into a more informative

asymmetry and a stronger training signal. In particular, we investigate which augmentations should be combined and how they should be composed. We utilize two of our best three-augmentation compositions: *composition A* draws one of mixed resolution, zoom-out onto a padded canvas, or downscaling with mild photometric jitter, each with probability 0.3, and identity with probability 0.1; *composition B* is identical composition except for a different canvas setup in mixed resolution.

Composition outperforms individual augmentations. Composing multiple augmentations outperforms the best single augmentation at both scales (77.28% vs. 76.81% at 4B and 78.47% vs. 77.84% at 9B; Figure 2c, left), and the gain rests on the structure rather than on any single member: removing the mixed-resolution or the zoom-out branch from composition A costs about one point (76.94% to 75.97% and 76.00%), and tilting the probabilities of composition B toward one augmentation also leads to a similar decrease (76.33% to 75.35%; Figure 2c, right).

Exclusive composition outperforms sequential composition. We compare *exclusive composition*, which samples and applies a single augmentation to each student view, with *sequential composition*, which stacks multiple augmentations on the same view (Figure 2d). Exclusive composition A achieves 75.89%, whereas sequentially applying two strong augmentations reduces performance to 74.36%; composition B shows the same trend (76.33% vs. 74.55%). A single strong augmentation already removes substantial visual detail, and stacking another can leave too little evidence for recovery, resulting in a large but less informative teacher–student gap.

Weak operators can be safely stacked onto strong ones. The opposite trend holds for weak augmentations (Figure 2e). Stacking hue/gamma jitter, rotation, blur, noise, or erasing onto a strong augmentation largely preserves performance, while color jitter improves it to 76.56% ($+0.7\%$). Unlike a second strong augmentation, a weak operator introduces appearance variation without substantially removing task-relevant evidence, providing additional regularization while preserving informative asymmetry. Our best recipe follows this pattern by stacking color jitter on the downscaling augmentation.

Best recipe. Our strongest configuration is an exclusive composition of three augmentations. With probability 0.9, one augmentation is sampled uniformly; otherwise the image is unchanged:

$$T(x) = A_k(x), \qquad k \sim \text{Categorical}(0.3, 0.3, 0.3, 0.1), \qquad A_4 = \text{identity}. \tag{6}$$

A_1 is *mixed resolution*, combining a low-resolution background with a high-resolution local patch; A_2 is *zoom-out*, shrinking the image to 0.4–$0.8\times$ and placing it on the original canvas; and A_3 is *downscaling with photometric jitter*, reducing resolution to 0.3–$0.6\times$ followed by random brightness, contrast, and color perturbations. These operators withhold fine-grained visual information in complementary ways, while A_1 and A_3 also reduce visual-token cost. We use this composition as the default T in S^{2}VOPD.

4.3 Main results

Comparison with open- and closed-source LLMs. Table 1 shows that S^{2}VOPD outperforms a majority of much larger VLMs, achieving an average accuracy of 78.85% at 9B. It surpasses all open-source models, including Qwen3.5-397B, as well as proprietary models such as GPT-5.4 and Gemini-3-Flash. These gains require neither ground-truth annotations, external rewards, nor a separately trained teacher, demonstrating that augmentation-induced teacher–student asymmetry alone provides an effective training signal.

Fair comparison with prior methods. To provide a fair competition, we train all several baseline methods in 4.1 under the same training context on Vision-OPD-6K and evaluation protocols. At both 4B and 9B scales, S^{2}VOPD achieves the highest average performance and leads on HR-4K, HR-8K, and MME-RealWorld, improving over the best self-rewarding baseline by 4.1% at 4B and 3.0% at 9B. More importantly, it also surpasses all three baselines with privileged information at both scales, despite using no annotations at all.

Generalization to math reasoning. To further prove the generalization of S^{2}VOPD, we evaluate on math reasoning, a domain out-of-distribution of the training set in Table 2. Although Vision-OPD-6K contains no math problems, S^{2}VOPD improves performance from 69.54% to 72.79% at 4B and from 71.99% to 75.97% at 9B, achieving the strongest overall results among the compared methods. In contrast, privileged-OPD and self-rewarding methods exhibit a cross-domain trade-off, improving either perception or mathematical reasoning at the expense of the other. These results suggest that

augmentation-induced discrepancy provides a more generalizable visual learning signal rather than rather than inducing specialization to a particular domain.

Table 1: Comparison with open-source and proprietary models, and prior models trained on Vision-OPD-6K, across six fine-grained perception benchmarks. **Bold** and underline mark the best and the second-best entry within each group (open-source, proprietary, Qwen3.5-4B, Qwen3.5-9B); § indicates that best instead of last checkpoint performance is reported.

Method	Perception						Avg
	V*	Zoom	HR-4K	HR-8K	MME-RW	MME-RW-CN	
Open-source models referencing (Wei et al., 2026) and (Yuan et al., 2026)							
MiniCPM-V-4.5	70.68	42.60	69.63	61.50	62.65	61.64	61.45
Qwen2.5-VL-7B	78.53	42.49	71.62	67.88	60.80	58.30	63.27
MiMo-VL-RL-7B	83.25	45.68	73.50	69.38	62.73	55.89	65.07
Qwen3-VL-4B	80.10	40.24	78.25	72.88	63.47	63.63	66.43
Thyme-7B	82.20	45.09	77.00	72.00	64.80	64.59	67.61
DeepEyesV2-7B	81.68	44.97	77.88	73.75	64.90	65.07	68.04
DeepEyes-7B	85.86	46.51	75.13	72.63	64.10	64.09	68.05
Qwen3-VL-8B-Instruct	84.82	42.96	79.63	75.25	63.19	64.61	68.41
GLM-4.5V	83.25	49.23	81.63	74.88	66.04	60.71	69.29
GLM-4.6V (106B)	86.91	50.06	82.13	78.88	65.57	65.62	71.53
Kimi-K2.6 (1T)	88.48	53.14	81.88	78.00	69.22	66.13	72.81
SenseNova-MARS-8B	**92.15**	47.81	83.13	78.38	67.90	68.90	73.05
Kimi-K2.5 (1T)	85.86	56.33	81.87	75.38	71.51	68.40	73.23
Qwen3-VL-235B-Instruct	91.10	56.09	86.13	80.38	71.74	69.04	75.75
Qwen3.5-397B	87.96	**57.16**	**89.38**	**85.50**	**74.82**	**69.82**	**77.44**
Proprietary models							
GPT-5.1	70.16	47.22	67.00	65.25	64.04	55.57	61.54
GPT-5.2	79.06	50.89	81.12	78.38	72.60	68.80	71.81
GPT-5.4	76.96	52.66	84.00	77.88	74.20	70.93	72.77
Gemini-3-Flash	86.39	59.29	87.88	85.00	74.86	72.62	77.67
Gemini-3.5-Flash	**89.01**	**61.42**	89.12	86.62	75.31	**73.97**	79.24
Gemini-3.1-Pro	87.96	61.18	**89.63**	**86.88**	**76.53**	73.31	**79.25**
Qwen3.5-4B							
Qwen3.5-4B	84.29	47.69	84.38	80.13	63.86	63.70	70.67
w/ privileged info.							
ZwZ-4B	**92.67**	55.74	81.75	79.50	68.52	68.09	74.38
OPSD	84.29	58.93	85.88	79.62	75.65	**73.43**	76.30
Vision-OPD-4B	91.10	**61.42**	82.12	80.38	74.38	69.98	76.56
w/o privileged info.							
Intuitor§	81.15	51.24	83.75	82.38	57.50	59.57	69.27
RENT§	80.63	51.01	82.88	81.62	58.55	60.93	69.27
TTRL§	85.96	55.62	83.00	81.88	69.85	65.56	73.64
S^2VOPD (ours)	91.10	57.75	**86.75**	**82.62**	**76.43**	71.91	**77.76**
Qwen3.5-9B							
Qwen3.5-9B	82.72	52.07	85.75	80.63	71.40	67.67	73.37
w/ privileged info.							
ZwZ-8B	**91.10**	58.11	84.38	82.00	69.87	70.59	76.01
OPSD	87.43	61.18	84.75	81.88	74.10	72.00	76.89
Vision-OPD-9B	87.43	**63.67**	86.12	85.12	70.01	68.89	76.87
w/o privileged info.							
RENT§	88.48	56.69	83.12	81.88	72.20	67.89	75.04
Intuitor§	89.01	56.88	85.75	83.50	72.18	67.91	75.87
TTRL§	87.96	56.14	84.62	81.50	72.90	68.08	75.20
S^2VOPD (ours)	90.58	59.76	**86.50**	**85.38**	**77.14**	**73.72**	**78.85**

4.4 Ablation Study and Analysis

We ablate the effectiveness of augmentation and EMA teacher in Table 3.

Asymmetry is necessary. Removing the student-view augmentation while keeping the objective (*w/o Aug.*, Table 3) collapses the gain back to the base level: 70.52% against the base model's 70.67% on the six perception benchmarks. With no information gap, the objective amplifies the teacher's confident errors instead of correcting the student's.

Table 2: Generalization to mathematical reasoning trained on perception data (Vision-OPD-6K). We adopt Qwen3.5-4B (left) and Qwen3.5-9B (right) as base models; AVG is over the three benchmarks. **Bold** and underline mark the best and the second-best; § indicates that best instead of last checkpoint performance is reported.

	Qwen3.5-4B				Qwen3.5-9B			
Method	**MathVista**	**MathVerse**	**MathVision**	**AVG**	**MathVista**	**MathVerse**	**MathVision**	**AVG**
Qwen3.5 (base)	75.80	67.56	65.26	69.54	78.80	70.25	66.91	71.99
w/ privileged info.								
ZwZ (4B / 8B)	71.80	40.48	52.17	54.82	76.00	56.45	57.01	63.15
OPSD	77.80	71.50	60.39	69.90	75.10	70.10	58.98	68.06
Vision-OPD	79.40	71.68	62.50	71.19	**82.30**	74.75	67.96	75.00
w/o privileged info.								
Intuitor§	**81.80**	71.98	65.23	73.00	80.50	75.56	69.57	75.21
RENT§	80.47	72.64	**66.74**	**73.28**	79.00	74.92	69.31	74.41
TTRL§	80.80	71.68	65.33	72.60	79.40	74.72	67.62	73.91
S²VOPD (ours)	79.70	**73.63**	65.03	72.79	82.10	**76.12**	**69.70**	**75.97**

Table 3: Ablation of S²VOPD on Qwen3.5-4B and different EMA decay rate. *w/o Aug.* removes the view asymmetry by showing the student the same clean image as the teacher. $\alpha = 0.95$ is our default.

Method	**MME-RW**	**MME-RW-CN**	**AVG**
S²VOPD	76.43	71.91	**77.76**
w/o Aug.	62.85	63.90	70.52
w/o EMA			
fixed teacher	**76.78**	**72.00**	76.61
$\alpha = 0.95$	76.43	71.91	**77.76**
$\alpha = 0.99$	76.70	71.81	76.84
$\alpha = 0.999$	76.38	71.56	76.81

EMA teacher. Freezing the teacher at the base model (*fixed teacher*) reaches 76.61%, 1.15% below the default and still 84% of the full gain over the base model. A teacher that never improves therefore supplies most of the supervision on its own, and the moving average adds roughly one point on top. Consistent with this, the loss from freezing is confined to the two benchmarks that reward locating a small target, V* and ZoomBench, while MME-RealWorld is unchanged (76.78% against 76.43%). Table 3 further varies the EMA decay from 0.95 to 0.999. The default 0.95 is the best setting at 77.76%; slowing the teacher costs 0.92% at 0.99 and 0.95% at 0.999, and the two slow settings are indistinguishable from each other (0.03% apart). A teacher that lags far behind the student therefore still supplies a usable reference distribution: the update rate modulates the learning signal rather than creating it.

EMA teacher. Freezing the teacher at the base model (*fixed teacher*) yields 76.61%, only 1.15% below the default while retaining 84% of the gain over the base model. Table 3 further varies the EMA decay rate from 0.95 to 0.999. The default 0.95 performs best at 77.76%; slower updates reduce the average by 0.92–0.95%. Thus, a fixed or slowly updated teacher remains effective, while an EMA teacher with relatively faster updates provides the better supervision.

S²VOPD is insensitive to the choice of divergence. Table 4 compares generalized Jensen–Shannon divergence (JSD) with its forward- and reverse-KL limits. Performance varies by only 0.55% across the three objectives: 77.76% with JSD, 77.68% with forward KL, and 77.21% with reverse KL. We adopt symmetric JSD by default, which balances the mass-covering behavior of forward KL and the mode-seeking behavior of reverse KL, providing a more balanced distillation signal under degraded student observations.

Table 4: Comparison of different divergences. AVG is over 6 perception benchmarks.

α	**MME-RW**	**MME-RW-CN**	**AVG**
0.0 (forward KL)	75.93	71.81	77.68
0.5 (JSD)	**76.43**	71.91	**77.76**
1.0 (reverse KL)	76.32	**72.59**	77.21

5 CONCLUSION

S²VOPD introduces a novel method to creates informative teacher–student asymmetry without any privileged annotations by withholding task-relevant visual information from the student. The teacher observes the original image, while the student is on-policy distilled from a degraded view, turning the model's own predictive discrepancy into a perception-aligned learning signal. Our analysis further shows that the effectiveness of S²VOPD depends on both the magnitude and semantics of the induced information gap: the augmentation must remove enough evidence to create a meaningful learning signal, while preserving sufficient task-relevant information for recovery to remain possible. Across benchmarks and model scales, S²VOPD consistently improves fine-grained visual perception while generalizing to mathematical reasoning. It substantially outperforms privileged OPD methods, which improve perception at the cost of reasoning, as well as self-rewarding RL methods, which better preserve reasoning but yield weaker perceptual gains.

AI USE STATEMENT

We used generative AI tools solely for language editing to improve the grammar, clarity, and readability of the manuscript. All AI-assisted edits were carefully reviewed to ensure that they preserved the intended meaning and technical accuracy. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work trains and evaluates on existing, publicly released benchmarks and models (Vision-OPD-6K, V*Bench, ZoomBench, HR-Bench 4K/8K, MME-RealWorld and its Chinese subset, MathVista, MathVerse, MathVision, and the Qwen3.5 model family); it introduces no new human-subject data collection, crowdsourcing, or personally identifiable data. The method itself only manipulates the visual input available to a student policy during training (through resolution, geometric, photometric, and occlusion transformations of the input image) and does not add capabilities beyond those already present in the base vision-language models used. We are not aware of ethical concerns beyond those already associated with the base models and benchmarks used.

REPRODUCIBILITY STATEMENT

The full training configuration – base models, batch size, rollout count, learning rate, EMA decay rate, prompt/response length budgets, and training data – is given in Section 4.1. The augmentation search space and the composition rule behind the default S^2VOPD recipe are specified in Section 3.2 and Appendix A. The evaluation protocol for all nine benchmarks –decoding settings, token budgets, and the answer-extraction/LLM-judge procedure – is likewise given in Section 4.1. We will release all training and evaluation code and deliver all model checkpoints for reproducibility.

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.

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.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

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.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15745–15753. IEEE, 2021.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 113–123. IEEE, 2019.

Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. *Advances in neural information processing systems*, 33:18613–18624, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar,

et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9726–9735. IEEE, 2020.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 103291–103325, 2026.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning, 2017. URL https://arxiv.org/abs/1610.02242.

Qiming Li, Tianlun Li, Xiaolong Cheng, Hangyu Li, Ruiyan Gong, Kangning Niu, Kaitao Jiang, and Mu Xu. Prpo: Perception-reinforced policy optimization via token-level dynamic advantage reshaping. *arXiv preprint arXiv:2606.08708*, 2026.

Yijiang Li, Xinjiang Wang, Lihe Yang, Litong Feng, Wayne Zhang, and Ying Gao. Diverse cotraining makes strong semi-supervised segmentor. *arXiv preprint arXiv:2308.09281*, 2023.

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.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation. *Advances in Neural Information Processing Systems*, 38:2923–2957, 2025.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. 2024. URL https://arxiv.org/abs/2310.02255.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *arXiv preprint arXiv:2602.04942*, 2026.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. Maximizing confidence alone improves reasoning. 2026. URL https://openreview.net/forum?id=Qhg479eBmo.

Hejian Sang, Yuanda Xu, Zhengze Zhou, Ran He, Zhipeng Wang, and Jiachen Sun. On-policy self-distillation for reasoning compression. *arXiv e-prints*, pp. arXiv–2603, 2026.

Yash Kumar Sharma, Charan Ramtej Kodi, and Vineet Padmanabhan. A self supervised learning framework for imbalanced medical imaging datasets. *arXiv preprint arXiv:2604.01947*, 2026.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Haoxiang Sun, Zhihang Yi, Langxuan Deng, Yuhao Zhou, Peiqi Jia, Jian Zhao, Li Yuan, Jiancheng Lv, and Tao Wang. V-zero: Answer-label-free on-policy distillation with contrastive evidence gating for fine-grained visual reasoning. *arXiv preprint arXiv:2606.25319*, 2026.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

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

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Yunjie Tian, Lingxi Xie, Jiemin Fang, Jianbin Jiao, and Qi Tian. Beyond masking: Demystifying token-based pre-training for vision transformers. *Pattern Recognition*, 162:111386, 2025.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=QWTCcxMpPA.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, Zhuosheng Zhang, and Weiran Huang. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

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

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020a.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695. IEEE, 2020b.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 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.

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.

Kawtar Zaher, Ilyass Moummad, Olivier Buisson, and Alexis Joly. Self-supervised learning as discrete communication. *arXiv preprint arXiv:2602.09764*, 2026.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VIII*, pp. 169–186, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73241-6. doi: 10.1007/978-3-031-73242-3_10. URL https://doi.org/10.1007/978-3-031-73242-3_10.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, volume 2025, pp. 89655–89701, 2025.

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

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=OU9nFEYR2M.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. TTRL: Test-time reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=VuVhgEiu20.

A Augmentation Operators

Table 5 lists the primitive operators of the augmentation space described in Section 3, grouped by family, together with the strength ranges used in our experiments.

Table 5: Overview of augmentation operators on the student policy.

Operator	Effects	Strengths Range
Information reduction		
Downscale	fewer pixels and visual tokens (no resize-back)	$s \sim \mathcal{U}(0.3, 0.6)$; 0.4–0.7; 0.25–0.5
Downscale, 3-tier	one of three bands drawn per sample	$[0.2, 0.35]$, $[0.35, 0.5]$, $[0.5, 0.75]$
Gaussian blur	soft low-pass	radius $\mathcal{U}(0.5, 1.5)$; $(1.5, 3.0)$; $(3.0, 6.0)$
Spectral band-stop	removes an annulus of the 2D spectrum	inner radius 0.15–0.35, width 0.15–0.3
Gaussian noise	additive; lowers signal-to-noise ratio	DDPM step 200 ($\sigma \approx 0.11$), ρ=0.5
Local pixelation	piecewise-constant low-pass (mosaic patches)	application probability ρ=0.4
Visual-token drop	fewer encoder tokens at native fidelity	drop $15/30/50\%$
Geometric		
Rotation	rotates the frame	$\pm 35^\circ$
Translation	shifts the frame	up to 15% of the side length
Crop	removes peripheral context, and sometimes the target	min. scale 0.7 / 0.5 / 0.3
Zoom-out	shrinks the scene onto a padded canvas	scale $\mathcal{U}(0.4, 0.8)$
Photometric		
Colour jitter	brightness / contrast / saturation, ρ=0.8 each	0.5–1.5 / 0.5–1.8 / 0.2–1.8
Heavy photometric	increase hue, gamma, sharpness, ρ=0.5 each	hue ± 0.15; gamma 0.6–1.6; sharp. 0.2–2.5
Occlusion		
Random erasing, grey fill	deletes patches, constant fill	$\leq$3 patches, 5–30% area
Random erasing, noise fill	deletes patches, random fill	$\leq$3 patches, 5–30% area
GridMask	regular-lattice masking, grey fill	cell period 0.1–0.3, keep ratio 0.5–0.7

B Full Related Work

Self- and semi-supervised learning.

A line of self-supervised representation learning derives supervision directly from augmented views of unlabeled images. Contrastive methods such as SimCLR (Chen et al., 2020) and MoCo (He et al., 2020) align different augmented views of the same image, while negative-free methods including BYOL (Grill et al., 2020), SimSiam (Chen & He, 2021), and DINO (Caron et al., 2021) match a student to a stop-gradient or momentum teacher under different augmentations. Similar asymmetry appears in semi-supervised consistency learning, where predictions from weakly augmented inputs supervise strongly augmented ones (Laine & Aila, 2017; Tarvainen & Valpola, 2017; Xie et al., 2020a; Sohn et al., 2020; Xie et al., 2020b; Li et al., 2023). Across these paradigms, augmentation is not merely regularization but a central source of supervision, and its design strongly shapes what is learned (Tian et al., 2020; Cubuk et al., 2019; 2020). We adopt this principle for token-level generative distillation: the teacher observes the clean image, the student an augmented view, and their next-token distribution gap provides the supervision for free.

Visual augmentation in vision-language models. Augmentation plays a much smaller role in vision-language model training than in representation learning, as geometric or photometric perturbations can alter instruction-relevant content such as text, object identity, and spatial relations. A recent exception is reinforcement learning for visual reasoning: NoisyRollout Liu et al. (2025) uses perturbed images to diversify exploration, while VPPO Huang et al. (2026) and PRPO Li et al. (2026) compare policy distributions across clean and perturbed views to identify perception-critical tokens and reweight policy gradients. In all cases, augmentation modulates learning driven by external rewards, with both views evaluated by the same policy. In contrast, we apply augmentation only to the student: the discrepancy from a teacher observing the clean image is itself the training signal, requiring neither rewards nor annotations.

Knowledge distillation and on-policy distillation. On-policy distillation (OPD) Agarwal et al. (2024) trains the student on its own trajectories while receiving dense token-level supervision from a teacher, reducing the train–inference mismatch of offline distillation. Subsequent work studies improved divergence objectives, selective supervision over informative tokens and trajectories Liu et al. (2026), and optimization strategies that better preserve visual grounding Yoon et al. (2026).

On-policy self-distillation (OPSD) extends this idea further and removes the need for a larger teacher by sharing a single model between student and teacher, with the teacher granted access to privileged information such as a ground-truth solution. Existing methods obtain this teacher-student asymmetry from verified reasoning traces Zhao et al. (2026a), additional context Ye et al. (2026), environmental feedback Hübotter et al. (2026), or privileged visual information Yuan et al. (2026); Liu et al. (2026); Yoon et al. (2026); Sun et al. (2026). These signals, however, require external annotation or guidance that might be costly or unavailable. We mitigate this by inverting where the asymmetry comes from. Rather than adding privileged information to the teacher, we subtract information from the student, by conditioning the student on a strongly augmented view of the same image. Our focus is therefore the design of this augmentation transformation and the properties that make this teacher-student discrepancy an effective supervisory signal.