VISUAL CONTRASTIVE SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) removes the external teacher required by on-policy distillation (OPD), but still needs asymmetric information between teacher and student so that the self-teacher provides a stronger learning signal. Existing methods create this asymmetry through privileged answers or visual evidence. We ask whether both can be removed, yielding a simpler form of OPSD driven purely by input conditioning. We propose **Visual Contrastive Self-Distillation**, namely **VCSD**, which converts image-content removal into an on-policy self-distillation signal. At each student-generated response prefix, the EMA teacher produces two next-token distributions under the same prompt and prefix: one conditioned on the original image and the other on a content-erased control. Their token-wise log-probability difference highlights candidates whose likelihood is specifically increased by the instance-level visual content. We use this contrast to sharpen the teacher's original-image distribution within its plausible support, and distill the resulting full-distribution target into the student. On ViRL39K, VCSD consistently outperforms matched OPSD across Qwen3-VL and Qwen3.5 models. It improves the seven-benchmark aggregate from 72.51% to 76.26% on Qwen3-VL-8B and from 74.97% to 79.24% on Qwen3.5-9B, with gains at all six evaluated scales. VCSD requires no external teacher, privileged answers, visual evidence signals, reasoning traces, or additional inference-time cost.

1 INTRODUCTION

On-policy distillation (OPD) (Agarwal et al., 2024; Lu & Lab, 2025; Li et al., 2026) trains a student on prefixes sampled from its own policy while using an external teacher for dense token-level supervision. This aligns training with the student's inference-time trajectories, but typically requires a stronger teacher, increasing post-training cost and complexity (Liu et al., 2026; Yoon et al., 2026). On-policy self-distillation (OPSD) removes the external teacher by deriving targets from the same model, often through an exponential moving average (EMA) teacher. However, when the student and self-teacher receive the same information at the same prefix, the target may add little beyond the student's current prediction. Thus, on-policy sampling determines *where* learning occurs, but not *what* additional supervision the self-teacher provides. Effective OPSD still requires a teacher–student asymmetry that makes the self-teacher more informative than the student.

Existing methods construct this asymmetry through auxiliary information that is available to the self-teacher but not to the student, as illustrated in Figure 1. For language reasoning tasks, the teacher may be conditioned on privileged answers or reasoning traces (Zhao et al., 2026), expert demonstrations (Shenfeld et al., 2026), or rich textual feedback (Hübotter et al., 2026). For vision-language tasks, the teacher can receive visual evidence signals, such as crops, regions, or other visual conditions that expose the relevant image evidence more directly (Yuan et al., 2026; Sun et al., 2026; Tian et al., 2026). These approaches can produce effective supervision, but depend on task-specific information or additional processing pipelines. These language-side signals are not always available, while visual evidence inputs can require annotations, localization models, or custom image transformations.

This motivates the central question of this work: ***Can the asymmetry required by on-policy self-distillation be constructed purely from input conditioning, without auxiliary information?*** We answer this question by treating a content-erased image not as an alternative teacher input to imitate, but as a controlled reference. At each student-generated response prefix, the same EMA teacher is evaluated twice under the same prompt and prefix: once with the original image and once with 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.

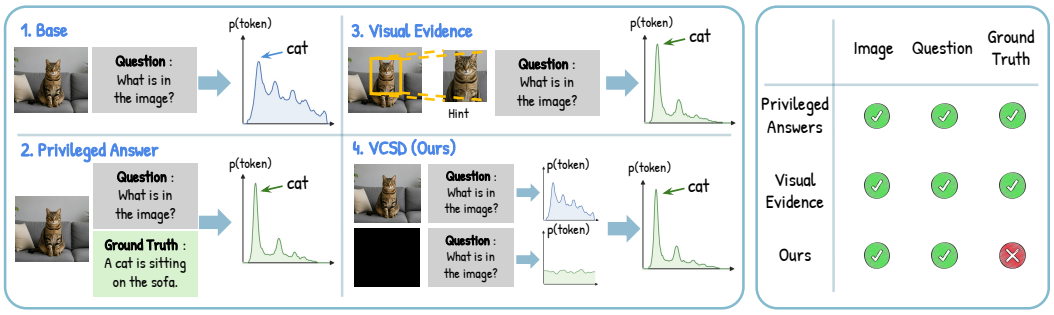


Figure 1: **Comparison of different sources of target asymmetry.** Existing methods sharpen the teacher distribution using privileged answers or visual evidence. Our method instead contrasts teacher predictions under the original image and a content-erased control, constructing a visually informed target using only the image and question.

content-erased control. Since only the visual condition changes, their token-wise log-probability difference captures how each candidate's likelihood changes when the instance-specific visual content is removed, as illustrated in Figure 1. This conditioning contrast provides the target asymmetry needed for self-distillation.

The contrast alone, however, does not define a reliable distillation target, since a token may exhibit a large relative change while remaining unlikely under the original image. We therefore assign complementary roles to the two teacher predictions: the original-image distribution identifies plausible candidates, while the conditioning contrast sharpens their relative preferences according to their dependence on visual content. Together, they yield a visually informed full-distribution target for distilling the student on its own trajectory.

Building on this idea, we propose **Visual Contrastive Self-Distillation**, namely VCSD. Given a student-generated prefix $y_{<t}$, we define

$$\begin{aligned} \Delta_t(v) &= \log p_\phi(v \mid P, J, y_{<t}) - \log p_\phi(v \mid P, J_{\text{ctrl}}, y_{<t}), \\ q_t^*(v) &\propto p_\phi(v \mid P, J, y_{<t}) \exp(\alpha \Delta_t(v)), \qquad v \in S_t, \end{aligned} \tag{1}$$

where S_t restricts the target to candidates that are plausible under the original image. The resulting full-distribution target is distilled into the student through forward KL along its on-policy trajectory. Here, "contrastive" refers to the contrast between conditional token distributions rather than an embedding-level contrastive loss.

We evaluate VCSD on ViRL39K across Qwen3-VL and Qwen3.5 models from 2B to 9B using seven vision-language benchmarks. It improves the aggregate score from 72.51% to 76.26% on Qwen3-VL-8B and from 74.97% to 79.24% on Qwen3.5-9B, with gains at every evaluated scale. It also consistently improves Qwen3.5 models by 2.8% to 4.3% over their corresponding base models. Across model families and scales, VCSD outperforms both the base models and matched OPSD baselines. Moreover, VCSD requires no external teacher, privileged answer, reasoning trace, evidence-focused crop, or external verifier.

Our contributions are threefold:

- We show that matched input conditioning can provide the target asymmetry required for OPSD without privileged answers or visual evidence signals.
- We propose VCSD, which uses the contrast between original-image and content-erased teacher predictions to sharpen a plausible full-distribution target.
- We demonstrate consistent improvements across Qwen3-VL and Qwen3.5 models from 2B to 9B on seven vision-language benchmarks.

2 Related Work

Target asymmetry in on-policy self-distillation. On-policy self-distillation builds its target from the same underlying model, requiring an asymmetry that makes it informative relative to the student. Ex-

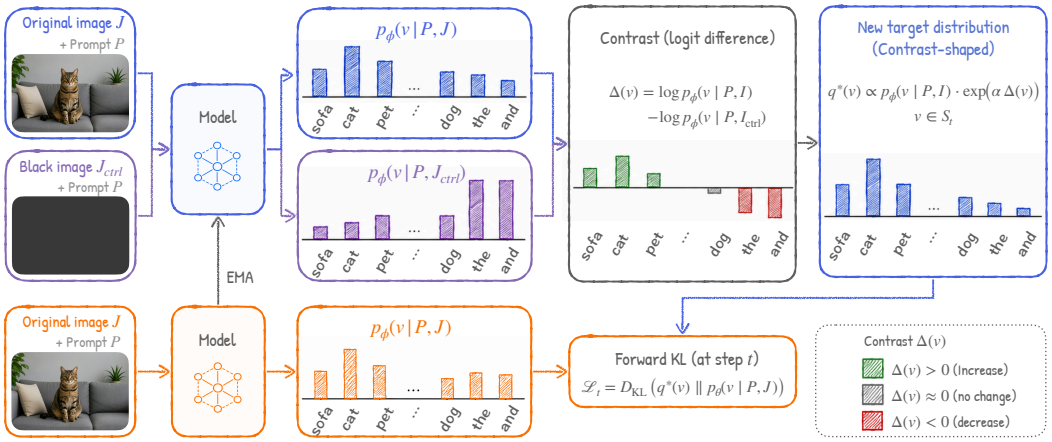


Figure 2: **Overview of VCSD.** At each student-generated prefix, the EMA teacher contrasts next-token distributions under the original image and a content-erased control. The contrast sharpens the plausible original-image distribution to form the target, which is distilled into the student via forward KL. The response prefix is omitted for clarity.

isting methods create this asymmetry with privileged answers or reasoning traces (Zhao et al., 2026), evidence-centered visual views (Yuan et al., 2026), or paired evidence for trajectory selection (Sun et al., 2026). Other supervision regimes use verifiable rewards (Shao et al., 2024; Yang et al., 2026) or separately trained teachers, sometimes with visual weighting or gradient steering (Liu et al., 2026; Bousselham et al., 2026; Yoon et al., 2026). VCSD instead derives target asymmetry from matched visual conditions of the same target model: the content-erased prediction is the reference, while the real-image prediction anchors the target to candidates supported by the input. This targets the tendency of VLMs to favor linguistic priors over fine-grained image evidence (Guan et al., 2024). At inference time, related methods address this dependency through contrastive decoding across models, visual conditions, or internal representations (Li et al., 2023; Leng et al., 2024), or through layer-wise activation refinement (Wang et al., 2025a). VCSD distills the visual contrast into OPSD without verifiable rewards, external teachers, or extra inference-time computation.

Relative predictive distributions under paired conditions. Differences between predictive distributions have been used to expose information that is not apparent from either prediction in isolation. Contrastive decoding compares expert and amateur models, or original and degraded visual conditions, to modify token selection at inference (Li et al., 2023; Leng et al., 2024). Beyond decoding, paired model behaviors can also define training signals: MARGO uses non-thinking rollouts as same-model references when estimating the advantage of explicit reasoning (Wang et al., 2026b). Related policy-transfer methods compare model checkpoints: weak-to-strong preference optimization transfers changes introduced by alignment (Zhu et al., 2025), while Direct-OPD transfers an RL-induced policy difference to a student's on-policy states (Feng et al., 2026). VCSD considers a same-model, paired-input setting in which the target parameters and response prefix are held fixed while only the visual condition changes. The resulting vocabulary-level contrast shapes a detached token distribution that is distilled by forward KL. The real-image prediction supplies the plausibility anchor, and the content-erased prediction supplies the reference used to adjust its relative token preferences.

3 Method

We propose VCSD, which constructs the target asymmetry required for OPSD from matched visual conditioning. As shown in Figure 2, the student generates an on-policy response under the original image, while the EMA teacher evaluates each response prefix under both the original image and a content-erased control. Their token-wise log-probability contrast sharpens the original-image distribution within its plausible support, yielding a full-distribution target for self-distillation.

3.1 On-Policy Self-Distillation Setup

Let $\mathcal{D} = \{(P_i, J_i)\}_{i=1}^{N}$ denote a collection of prompt–image pairs, where P is the prompt and J is the original image. We denote the trainable student by π_θ and its EMA teacher by π_ϕ. The teacher serves as a stop-gradient target; its parameters receive no gradients.

For each $(P, J) \sim \mathcal{D}$, the student samples an on-policy response

$$y \sim \pi_\theta(\cdot \mid P, J). \tag{2}$$

At generation step t, both the student and teacher are evaluated along the same student-generated prefix $y_{<t}$. The student next-token distribution is

$$p_{\theta,t}(v) = \pi_\theta(v \mid P, J, y_{<t}), \qquad v \in \mathcal{V}, \tag{3}$$

where $\mathcal{V}$ is the vocabulary.

The rollout is on-policy because the student determines every prefix at which supervision is applied. However, on-policy sampling alone does not specify how the teacher target should become more informative than $p_{\theta,t}$. Instead of providing the teacher with a privileged answer or an evidence-focused crop, VCSD constructs this target asymmetry by comparing the teacher under two matched visual conditions.

3.2 Visual Conditioning Contrast

We construct a content-erased control $J_{\mathrm{ctrl}} = \mathcal{C}(J)$, where $\mathcal{C}(J)$ is a same-size black RGB image. The control preserves the image resolution, multimodal input interface, preprocessing path, and visual-token count, while removing the instance-specific content of J.

At each fixed prefix $y_{<t}$, the EMA teacher produces two next-token distributions:

$$p^{J}_{\phi,t}(v) = \pi_\phi(v \mid P, J, y_{<t}), \quad p^{0}_{\phi,t}(v) = \pi_\phi(v \mid P, J_{\mathrm{ctrl}}, y_{<t}). \tag{4}$$

The prompt, response prefix, model parameters, and multimodal computation path are shared between the two evaluations. They differ only in whether the instance-specific visual content is present.

We measure the change in token preference through the vocabulary-level log-probability contrast

$$\Delta_t(v) = \log p^{J}_{\phi,t}(v) - \log p^{0}_{\phi,t}(v). \tag{5}$$

A positive $\Delta_t(v)$ indicates that token v receives greater support under the original image than under the content-erased control. Conversely, a negative value indicates that its probability increases when the original visual content is removed. Thus, Δ_t describes how the teacher's next-token preferences change in response to the instance-specific visual content.

Importantly, the control distribution is used only as a reference. It is neither treated as a teacher target nor directly distilled into the student.

3.3 Contrast-Shaped Teacher Target

The conditioning contrast provides a direction for modifying the teacher target, but it is not sufficient on its own. A token may exhibit a large relative change between the two conditions while remaining highly unlikely under the original image. We therefore use the original-image teacher distribution as a plausibility anchor and apply contrastive shaping only to candidates that it considers sufficiently likely.

Specifically, we define the relative plausibility support

$$\mathcal{S}_t(\beta) = \left\{ v \in \mathcal{V} \;:\; p^{J}_{\phi,t}(v) \geq \beta \max_{u \in \mathcal{V}} p^{J}_{\phi,t}(u) \right\}, \tag{6}$$

where $\beta \in [0, 1]$ controls the support threshold. This relative-threshold construction follows the plausibility restriction commonly used in contrastive decoding (Li et al., 2023; Leng et al., 2024).

We then construct the contrast-shaped teacher target as

$$q_t^\star(v) = \frac{\mathbf{1}\left[v \in \mathcal{S}_t(\beta)\right] p_{\phi,t}^J(v) \exp\Big(\alpha \widetilde{\Delta}_t(v)\Big)}{\sum_{u \in \mathcal{S}_t(\beta)} p_{\phi,t}^J(u) \exp\Big(\alpha \widetilde{\Delta}_t(u)\Big)}, \tag{7}$$

where $\alpha \geq 0$ controls the strength of contrastive shaping. Here, $\widetilde{\Delta}_t$ equals Δ_t except that the contrast values of designated sequence-termination tokens are set to zero.

Equation 7 assigns complementary roles to the two signals. The original-image distribution $p_{\phi,t}^J$ determines the initial probability and the admissible candidate set, while $\widetilde{\Delta}_t$ adjusts the relative probabilities according to how strongly each candidate is favored by the original visual content. Within the support, the corresponding unnormalized log score is

$$\log p_{\phi,t}^J(v) + \alpha \widetilde{\Delta}_t(v) = (1+\alpha) \log p_{\phi,t}^J(v) - \alpha \log p_{\phi,t}^0(v), \tag{8}$$

for non-termination tokens. When $\alpha = 0$, the target reduces to the original-image teacher distribution renormalized over $\mathcal{S}_t(\beta)$. When $\beta = 0$, the target spans the full vocabulary.

In VCSD, "contrastive" therefore refers to the contrast between two conditional token distributions, rather than positive and negative embedding pairs.

3.4 On-Policy Distillation and EMA Update

The contrast-shaped target is distilled into the student at every position of the student-generated response using full-distribution forward KL:

$$\mathcal{L}_{VCSD} = T_{\mathrm{KD}}^2 \, \mathbb{E}_{\substack{(P,J)\sim\mathcal{D} \\ y\sim\pi_\theta(\cdot|P,J)}} \left[\frac{1}{|y|} \sum_{t=1}^{|y|} D_{\mathrm{KL}}\left(\mathrm{sg}\left[q_t^\star\right] \,\|\, p_{\theta,t}\right) \right], \tag{9}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient. All teacher and student distributions used for target construction and distillation are evaluated with the same distillation temperature T_{KD}; its dependence is omitted from the notation for clarity. The factor T_{KD}^2 follows standard temperature-scaled knowledge distillation.

Gradients pass only through the student distribution $p_{\theta,t}$. The two teacher evaluations are performed on the same fixed student-generated prefix and do not produce separate response trajectories. After each student update, the teacher parameters are updated by $\phi \leftarrow \mu\phi + (1-\mu)\theta$, where $\mu \in [0, 1)$ is the EMA decay coefficient.

Training requires only the original prompt–image pair and the student's own on-policy response. It does not consume an external teacher, privileged answer, reasoning trace, evidence-focused crop, external verifier, or verifiable reward. At inference time, only the updated student is retained; neither the EMA teacher nor the content-erased control introduces an additional inference-time branch.

3.5 Policy-Improvement Interpretation

The conditional log-ratio in Equation 5 acts as a dense token-level reward for support attributable to instance-specific visual content. Under this view, the contrast-shaped target is the unique optimizer of a one-step KL-regularized policy-improvement objective: the original-image prediction supplies the reference policy, the conditioning contrast supplies the reward, and the plausibility set constrains the update to candidates already supported by the observed input. This connects input-conditioned self-distillation to conservative policy improvement without requiring an external reward model; Appendix C gives the complete formulation and its conditional-information interpretation.

4 Experiments

4.1 Setup

Models and training data. We evaluate VCSD on Qwen3-VL (Bai et al., 2025) at 2B / 4B / 8B and on Qwen3.5 (Qwen Team, 2026) at 2B / 4B / 9B. The main experiments train all post-training methods

Table 1: **Main results on seven benchmarks.** Base, SFT, GRPO, answer-hint OPSD, and VCSD are evaluated across three Qwen3-VL and three Qwen3.5 scales. Acc. is the unweighted mean of the seven benchmarks. VCSD provides a strong reward-free result and improves the corresponding base model at every scale. The highest score in each column within each model block is shown in bold.

Model	Method	BLINK	MMStar	V*	MathVista	HR4K	HR8K	HalluB	Acc.
Qwen3-VL-2B	Base	54.55	57.00	73.82	62.40	71.12	67.62	52.32	62.69
	SFT	50.92	58.40	70.68	61.30	71.62	68.50	50.63	61.72
	GRPO	55.29	62.13	74.87	64.90	76.62	71.50	**53.84**	65.59
	OPSD	53.87	60.67	73.30	63.20	72.88	71.00	50.53	63.63
	VCSD (ours)	**57.92**	**63.67**	**78.53**	**68.50**	**77.62**	**72.75**	53.41	**67.49** (+4.80)
Qwen3-VL-4B	Base	**67.18**	68.93	80.63	73.90	79.88	73.75	54.81	71.30
	SFT	66.02	68.60	82.72	75.10	78.62	74.12	52.26	71.06
	GRPO	65.70	**69.80**	**83.77**	**75.60**	**81.12**	71.37	60.14	72.50
	OPSD	64.97	68.73	**83.77**	74.10	78.12	75.63	54.50	71.40
	VCSD (ours)	66.86	68.73	**83.77**	74.90	80.50	**77.00**	**60.37**	**73.16** (+1.86)
Qwen3-VL-8B	Base	69.65	70.67	82.72	76.50	77.38	71.12	59.51	72.51
	SFT	66.33	72.00	86.39	75.80	79.50	75.62	60.27	73.70
	GRPO	68.65	73.40	86.39	**79.30**	78.75	74.87	**60.83**	74.60
	OPSD	67.81	70.20	85.34	77.30	81.12	74.62	59.64	73.72
	VCSD (ours)	**70.38**	**74.07**	**87.43**	77.90	**84.12**	**79.25**	60.69	**76.26** (+3.75)
Qwen3.5-2B	Base	59.02	67.67	83.25	74.10	73.75	72.25	50.22	68.61
	SFT	59.65	67.87	79.06	73.00	76.25	72.25	52.67	68.68
	GRPO	64.07	66.47	**86.39**	72.80	76.38	73.50	51.28	70.13
	OPSD	58.76	66.40	79.06	71.00	76.62	75.88	47.89	67.94
	VCSD (ours)	**66.44**	**69.53**	83.25	**75.90**	**78.75**	**77.50**	**56.47**	**72.55** (+3.94)
Qwen3.5-4B	Base	64.97	73.00	82.20	81.70	81.38	74.00	60.36	73.94
	SFT	65.39	73.00	86.39	81.30	83.50	77.50	61.66	75.54
	GRPO	65.75	73.60	**86.91**	79.10	**86.88**	73.87	58.67	74.97
	OPSD	64.44	71.40	84.29	77.60	82.75	78.75	58.22	73.92
	VCSD (ours)	**66.70**	**75.00**	84.82	**82.20**	85.38	**79.38**	**63.95**	**76.77** (+2.83)
Qwen3.5-9B	Base	66.70	74.60	83.25	81.60	82.38	77.38	58.89	74.97
	SFT	67.02	75.60	84.29	81.60	82.12	78.25	60.56	75.63
	GRPO	68.86	75.93	86.39	81.50	81.62	79.38	60.97	76.38
	OPSD	67.65	72.27	**87.96**	78.80	81.50	78.12	56.80	74.73
	VCSD (ours)	**72.23**	**78.87**	85.86	**85.00**	**87.00**	**81.50**	**64.19**	**79.24** (+4.27)

on ViRL39K single-image dataset (Wang et al., 2026a). Section 4.3 additionally retrains VCSD on Vision-OPD6K training set (Yuan et al., 2026) to evaluate generalization across training data.

Baselines. Each model block compares the unmodified base model, the published answer-hint OPSD method (Zhao et al., 2026), and VCSD. GRPO uses rule-based verifiable rewards and its own tuned hyperparameters, while SFT uses cached non-thinking teacher responses. All arms match optimization steps and prompt budget. The OPSD target model is conditioned on the reference answer while the student receives only the original question and image. VCSD retains the on-policy rollout structure but constructs its target from the paired original and content-erased image conditions without consuming the answer. Implementation details and training hyperparameters are provided in Appendix A.

Benchmarks and aggregate metric. We evaluate general visual perception with BLINK (Fu et al., 2024) and MMStar (Chen et al., 2024), visual mathematics with MathVista (Lu et al., 2024), fine-grained and high-resolution perception with V*Bench (Wu & Xie, 2023) and HRBench4K/8K (Wang et al., 2025b), and hallucination with HallusionBench (Guan et al., 2024). We use GPT-5.4-mini for model-based judging. The reported *Acc* is the unweighted mean over these seven benchmarks.

4.2 Main Results

Table 1 consolidates the primary comparison across various Qwen model families and scales.

A single reward-free objective delivers gains across all six scales. VCSD raises the corresponding base-model aggregate at every scale from Qwen3-VL-2B/4B/8B to Qwen3.5-2B/4B/9B, with gains between $+1.86\%$ and $+4.80\%$. These results use neither cached target responses nor a verifiable reward, isolating input-conditioned self-distillation as the training signal.

VCSD improves over answer-hint OPSD across visual capabilities. On Qwen3-VL-2B, VCSD improves every benchmark over OPSD. This yields a $+3.86\%$ aggregate gain, with improvements

Table 2: **Generalization to Vision-OPD6K.** VCSD uses image–question pairs only; Vision-OPD rows use released checkpoints trained on the full corpus. Acc. averages the seven shared benchmarks, Avg. (9) adds MME-RealWorld CN/EN, and bold marks the highest score per model block.

Model	Method	BLINK	MMStar	V*	MathVista	HR4K	HR8K	HalluB	Acc.	MME-RW CN	MME-RW EN	Avg. (9)
Qwen3.5-4B	Base	64.97	73.00	82.20	81.70	81.38	74.00	60.36	73.94	64.20	62.90	71.63
	Vision-OPD	60.97	71.47	**87.43**	79.10	82.62	79.12	58.89	74.23	**69.21**	**70.44**	73.25
	VCSD (ours)	**67.65**	**74.73**	84.29	**82.70**	**85.88**	**81.12**	**62.35**	**76.96**	68.62	69.51	**75.21**
Qwen3.5-9B	Base	66.70	74.60	83.25	81.60	82.38	77.38	58.89	74.97	67.47	69.58	73.54
	Vision-OPD	66.23	75.33	**92.15**	82.60	85.50	**85.38**	59.44	78.09	**71.17**	72.20	76.67
	VCSD (ours)	**69.81**	**77.47**	90.58	**83.50**	**88.25**	83.62	**61.57**	**79.26**	68.35	**72.54**	**77.30**

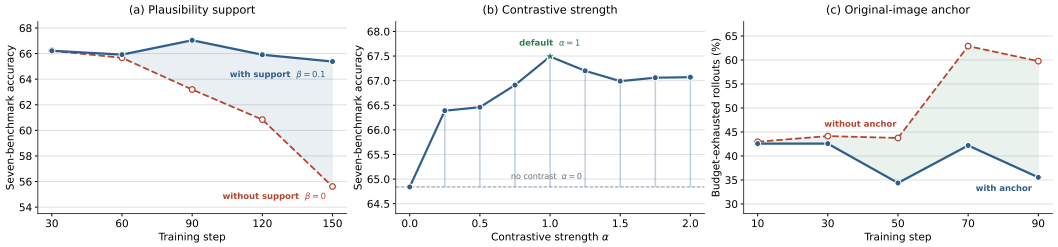


Figure 3: **Target-shaping ablations on Qwen3-VL-2B. (a)** Seven-benchmark accuracy over extended training with and without plausibility support: the unrestricted target ($\beta = 0$) degrades under continued self-distillation, while $\beta = 0.1$ preserves accuracy. **(b)** Accuracy across contrast strengths: accuracy rises monotonically to its peak at $\alpha = 1$ and stays within 0.6 points of the peak for every strength in $[0.75, 2.0]$; $\alpha = 1$ is the default. **(c)** Fraction of training rollouts that exhaust the generation budget without producing an answer, in a paired run differing only in the anchor coefficient: removing the anchor roughly doubles this fraction by the end of training.

spanning general perception, visual mathematics, fine-grained and high-resolution perception, and hallucination rather than being driven by a single benchmark. Since OPSD supplies the teacher with the reference answer, this comparison supports visual conditioning as an alternative source of informative targets across these capabilities.

Input-conditioned asymmetry scales across both model families. The gains over answer-hint OPSD persist at the largest tested scale in each family. This extends the result from smaller models to stronger starting models, showing that the contrast remains useful as base capabilities increase within the tested families. Matched visual predictions therefore provide effective self-distillation without answer or evidence hints.

4.3 Generalization Across Training Datasets

The main results train VCSD on ViRL39K. To test transfer across training corpora, we retrain VCSD on Vision-OPD6K using only its image–question pairs, without consuming bounding-box annotations. Table 2 compares Qwen3.5-4B and Qwen3.5-9B with Vision-OPD.

VCSD transfers beyond ViRL39K without supervision. In Table 2, VCSD reaches 76.96% and 79.26% aggregate accuracy on Qwen3.5-4B and Qwen3.5-9B, improving their corresponding base models by 3.02% and 4.29%. At both scales, VCSD also exceeds the released Vision-OPD on both Acc. and Avg. (9), using only image–question pairs from the 6K subset, while Vision-OPD uses bounding-box supervision on the full corpus. These results establish transfer across training corpora and higher aggregate performance without external information.

4.4 Effect of Plausibility Restriction

Plausibility support protects accuracy during continued training. To examine its role, we compare training with plausibility support enabled ($\beta = 0.1$) and removed ($\beta = 0$). As shown in Figure 3(a), the two runs are indistinguishable through step 60, after which the unsupported run degrades steadily, dropping to 55.6% by step 150, while the supported run stays near its starting accuracy (65–67%). The resulting gap is 3.8 points at step 90 and 9.8 points at step 150. Without this restriction, contrast can promote tokens that are unlikely under the original image simply because they are even less

Table 3: **Original-image anchor ablation on Qwen3-VL-2B.** No anchor removes the original-image log-probability term but retains contrast and plausibility support. Acc. averages seven benchmarks. Accuracy is comparable; Figure 3(c) shows the rollout-completion benefit.

Config	BLINK	MMStar	V*	MathVista	HR4K	HR8K	HalluB	Acc.
No anchor	58.71	62.93	75.92	67.00	76.38	73.12	53.43	66.78
With anchor (VCSD)	57.92	63.67	78.53	68.50	77.62	72.75	53.41	67.49

Table 4: **Distillation-divergence ablation on Qwen3-VL-2B.** Forward KL and JSD preserve the contrast-shaped target more effectively than reverse KL.

Divergence	BLINK	MMStar	V*	MathVista	HR4K	HR8K	HalluB	Acc.
forward KL	**57.92**	**63.67**	**78.53**	68.50	**77.62**	72.75	53.41	**67.49**
JSD	56.92	61.73	**78.53**	**69.20**	76.50	**73.75**	**54.61**	67.32
reverse KL	57.50	60.13	75.92	62.60	73.12	70.25	51.47	64.43

likely under the erased image. Plausibility support excludes these candidates from the training target, keeping repeated self-distillation focused on plausible predictions.

4.5 Effect of Contrastive Strength

VCSD is effective across a broad range of contrast strengths. To assess sensitivity to contrast strength, we vary α while keeping all other settings fixed. The $\alpha = 0$ control retains original-image self-distillation and plausibility support but removes contrastive reweighting. As shown in Figure 3(b), accuracy increases from 64.84% without contrast to 67.49% at $\alpha = 1$; even the smallest tested strength ($\alpha = 0.25$) recovers 1.55 points, and all tested strengths in $[0.75, 2.0]$ remain within 0.58% of the peak. This broad effective range indicates that the benefit comes from incorporating visual contrast rather than precisely tuning its strength. We use $\alpha = 1$ throughout the main experiments.

The original-image anchor improves rollout completion. We examine how the original-image anchor affects answer completion during training. In paired runs, we remove the standalone original-image log-probability term while retaining visual contrast and plausibility support. As shown in Figure 3(c), removing the anchor roughly doubles the fraction of rollouts that exhaust the generation budget without producing an answer by the end of training. Table 3 reports comparable seven-benchmark accuracy. The anchor therefore complements plausibility restriction: the mask selects admissible tokens, while retaining original-image preferences helps rollouts complete their answers.

4.6 Effect of Distillation Divergence

Forward KL directly transfers the target's visual preferences. To isolate how the shaped target supervises the student, we keep target construction unchanged and vary the distillation divergence. Table 4 shows comparable accuracy for forward KL and JSD. Forward KL exceeds reverse KL by 3.06% in aggregate accuracy, while differing from JSD by 0.17%. Forward KL directly reinforces tokens promoted by the visual contrast, including those initially assigned low student probability. Reverse KL weights discrepancies by the student's current probabilities, which can weaken this correction; this distinction motivates forward KL as our default.

Content removal is the key property of the reference. To determine what makes a reference effective, we compare black, Gaussian-noise, blurred, and no-image controls on Qwen3-VL-2B. These variants distinguish removing instance-specific content from merely degrading image appearance. Table 5 shows that black, noise, and no-image controls all outperform blur, with black improving aggregate accuracy by 3.49% and MathVista by 6.70% over blur. We additionally replace black controls with shuffled images drawn from other training samples, keeping the training recipe unchanged. This comparisons yield comparable aggregate accuracy with natural-image and black controls, suggesting that the benefit does not depend on a black reference specifically. Together, these results support removing the current instance's evidence from the reference rather than retaining a coarse view of the same scene. We therefore use black images as a simple deterministic reference throughout the main experiments.

Table 5: **Control-image construction on Qwen3-VL-2B.** Full-removal controls behave similarly, while blur preserves coarse image content and trails. Natural references use other samples' images.

Config	BLINK	MMStar	V*	MathVista	HR4K	HR8K	HalluB	Acc.
Gaussian noise	**58.50**	**63.80**	76.44	64.50	76.63	**73.12**	53.04	66.58
Black (ours)	57.92	63.67	**78.53**	**68.50**	**77.62**	72.75	53.41	**67.49**
Unrelated natural image	56.86	63.60	75.92	66.20	77.50	72.62	53.92	66.66
Gaussian blur	55.18	60.87	75.92	61.80	74.87	69.38	49.98	64.00
No image	56.65	62.73	76.44	62.70	76.50	71.75	**53.95**	65.82

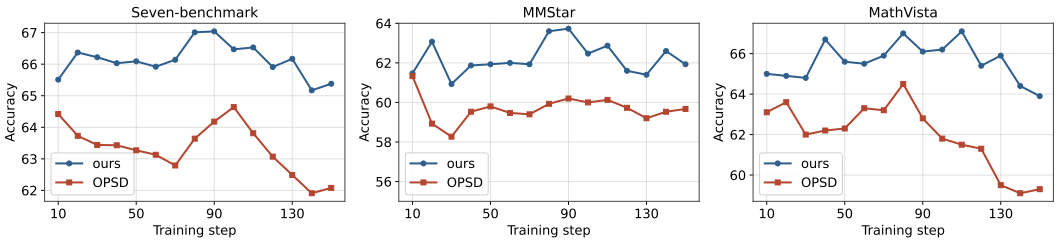


Figure 4: **Training dynamics of VCSD and OPSD on Qwen3-VL-2B.** Each panel tracks accuracy over training for the two runs: seven-benchmark aggregate, MMStar, and MathVista. VCSD leads on all three metrics throughout the evaluated training trajectory.

VCSD maintains an accuracy advantage over OPSD throughout training. To examine whether the gains persist as training proceeds, we compare VCSD and answer-hint OPSD at matched training steps on Qwen3-VL-2B. We track seven-benchmark mean accuracy, MMStar, and MathVista to examine aggregate performance, general perception, and visual mathematics. As shown in Figure 4, VCSD consistently leads on the aggregate, with an average absolute improvement of 2.80% across the evaluated steps. The advantage also holds on MMStar and MathVista throughout the measured trajectory, showing that visual contrast provides an accuracy advantage during both early learning and continued optimization.

VCSD benefits from visual-contrast training and teacher updates. We investigate how visual-contrast targets and teacher updates contribute to VCSD's gains. On Qwen3-VL-2B, we compare three controls: inference-only contrast applies the same contrast and plausibility mask at decoding without training; sharpening-only training doubles the original-image logits while removing the control branch; and the frozen-teacher variant fixes teacher parameters while retaining contrast and student training. As shown in Figure 5, VCSD improves accuracy by 4.80% over the base, compared with -0.20% for inference-only contrast; both training controls also yield smaller gains. These results support learning visually directed target changes beyond generic sharpening, with teacher updates contributing to the improvement.

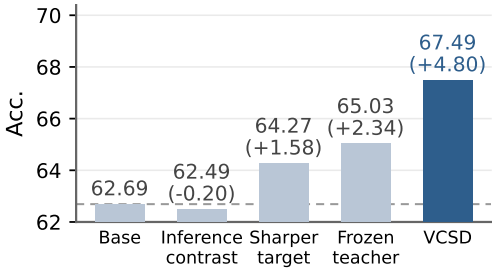


Figure 5: **Sources of VCSD's gains on Qwen3-VL-2B.** Bars report seven-benchmark mean accuracy. Parentheses show absolute changes from the base model, whose accuracy is marked by the dashed line.

5 Conclusion

We introduced VCSD, which constructs OPSD target asymmetry from matched original and content-erased image conditions. Their token-wise log-probability contrast sharpens the teacher's original-image distribution within plausible support, producing a full-distribution target without an external teacher, privileged answers, or external visual-evidence annotations. Across six Qwen3-VL and Qwen3.5 scales, VCSD improves the corresponding base models and matched OPSD baselines on ViRL39K, with gains over the base models transferring to Vision-OPD6K. Together, the consistent gains and target-design analyses establish conditional prediction changes as effective self-distillation supervision. At deployment, resulting model retains backbone's standard single-pass inference path.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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

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.

Shiyuan Feng, Huan-ang Gao, Haohan Chi, Hanlin Wu, Zhilong Zhang, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Weak-to-strong generalization via direct on-policy distillation. *arXiv preprint arXiv:2607.05394*, 2026.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024.

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.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 12286–12312, 2023.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan-ang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 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.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

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. In *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

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.

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

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, 2026a.

Kaishen Wang, Hengrui Gu, Meijun Gao, and Kaixiong Zhou. Damo: Decoding by accumulating activations momentum for mitigating hallucinations in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Kaishen Wang, Tong Zheng, Xuehao Cui, Ruibo Chen, Tianyi Xiong, and Heng Huang. Mitigating factual hallucination in large reasoning models via mixed-mode advantage regularization. *arXiv preprint arXiv:2607.05861*, 2026b.

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, 2025b.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 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.

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

Wenhong Zhu, Zhiwei He, Xiaofeng Wang, Pengfei Liu, and Rui Wang. Weak-to-strong preference optimization: Stealing reward from weak aligned model. In *International Conference on Learning Representations*, volume 2025, pp. 37116–37144, 2025.

APPENDIX

A EXPERIMENTAL DETAILS

A.1 BASELINE IMPLEMENTATIONS

GRPO. The reward is a rule-based verifiable check with no reward model and no KL term in the reward: the predicted answer is extracted from the response and matched against the reference. Rollouts sample $n = 8$ responses per prompt at temperature 1.0 with a maximum response length of 1024. GRPO uses its own tuned hyperparameters at every scale (learning rate 1×10^{-6}, PPO mini-batch 8, clip range 0.2–0.28, 32 prompts per step) with no per-scale hyperparameter search.

SFT. The teacher cache is generated by the same backbone at each scale, sampling the fixed-seed $6{,}000$-prompt subset at temperature 1.0 under the non-thinking chat template, whose generation prompt closes the reasoning block so that targets are direct answers.

VCSD. Unless stated otherwise: $\alpha = 1$, $\beta = 0.1$, $T_{\mathrm{KD}} = 2$, EMA rate $\rho = 0.05$, uniform response-position weights, learning rate 2×10^{-6}, batch 32 prompts with $n = 8$ rollouts, 10 warmup steps, and a fixed 90-step budget; the full-vocabulary forward-KL objective is optimized with AdamW. Every training run in the paper uses a single 8-GPU node. Measured wall clock for a 90-step VCSD run on Qwen3-VL-2B is about 2.0 hours.

A.2 VARIANCE ACROSS INDEPENDENT TRAINING SEEDS

Table 6 reports all settings trained with multiple independent seeds. Every checkpoint is evaluated under the same protocol, and VCSD outperforms GRPO, which in turn outperforms OPSD, for every seed.

Table 6: **Variance across independent training seeds on Qwen3-VL-2B.** Each row reports the mean and standard deviation of the seven-benchmark accuracy over three independent training seeds.

VCSD	GRPO	OPSD
67.49 ± 0.51	65.69 ± 0.06	63.63 ± 0.77

B SUPPLEMENTARY EXPERIMENTS AND ANALYSES

B.1 VISION-OPD NATIVE EVALUATION

Native evaluation separates localization and high-resolution strengths. On Qwen3.5-4B, VCSD without boxes provides the strongest HRBench4K and MME-RealWorld EN results, while the released Vision-OPD checkpoint leads on V* and MME-RealWorld CN. Composing input contrast with the bbox-crop teacher reaches the strongest ZoomBench, HRBench8K, and native average, identifying the two supervision signals as complementary. ZoomBench is measured under the separate native protocol of Table 7 and is not included in Acc. A matched no-contrast control (the Vision-OPD recipe retrained on the same data for the same 65 steps) scores 90.58/80.38/79.00/60.83 on V*/HRBench4K/HRBench8K/ZoomBench under the native protocol; adding the contrastive term contributes $+4.12$ on HRBench4K, $+3.12$ on HRBench8K, and $+1.42$ on ZoomBench, while retaining a V* score of 89.53.

B.2 EFFECT OF VOCABULARY SUPPORT

The objective is insensitive to the distillation support. We compare the full-vocabulary objective against variants that restrict the support at each position to the top-K tokens of the student, or of the union of student and target, renormalizing both distributions on that set. At $K{=}10$ the union variant matches full-vocabulary distillation across five backbones: 73.64 versus 73.16 on Qwen3-VL-4B and 76.91 versus 76.77 on Qwen3.5-4B (ViRL39K), 76.84 versus 76.96 on Qwen3.5-4B and 78.44 versus 79.26 on Qwen3.5-9B (Vision-OPD6K), 74.80 versus 74.27 on Qwen3-VL-8B (Vision-OPD6K),

Table 7: **Vision-OPD native evaluation on Qwen3.5-4B.** All rows use non-thinking inference and the official `gpt-oss-120b` judge. VCSD and its bbox-crop variant train on Vision-OPD6K; the released Vision-OPD checkpoint uses the full corpus. Sup. denotes bounding-box supervision: the combined variant uses annotated crops for the teacher view together with visual contrast. Avg. is the unweighted mean of the six displayed benchmarks, and bold marks column maxima. The combined variant has the highest mean, supporting the complementarity of crop supervision and contrast.

Model	Method	Sup.	V*	ZoomBench	HR8K	HR4K	MME-RW CN	MME-RW EN	Avg.
Qwen3.5-4B	Base	–	81.68	48.52	80.38	84.25	62.57	63.31	70.12
	Vision-OPD (official)	bbox	**90.05**	60.59	79.88	82.75	**69.26**	71.32	75.64
	VCSD (ours)	none	85.34	54.44	80.38	**85.62**	68.35	**73.40**	74.59
	VCSD + bbox-crop teacher	bbox	89.53	**62.25**	**82.12**	84.50	68.36	72.02	**76.46**

Table 8: **Vocabulary support for the distillation loss.** We compare full-vocabulary distillation with student top-K support and the union of the student's and target's top-K candidates, renormalizing both distributions on the selected set. Columns correspond to Qwen3.5-4B trained on Vision-OPD6K and Qwen3-VL-4B trained on ViRL39K, each evaluated in one batch. Scores average the six main benchmarks excluding HRBench8K; this mean is specific to this table. Bold marks the best score per column. Union-support variants remain close to full-vocabulary distillation, with no shared best K across the two settings.

Support	Qwen3.5-4B	Qwen3-VL-4B
full vocabulary	76.27	72.56
student top-10	75.35	72.66
student ∪ target top-10	76.24	**73.23**
student ∪ target top-50	**76.94**	72.53
student ∪ target top-100	75.89	72.95

and 65.98 versus 67.49 on Qwen3-VL-2B (ViRL39K). Table 8 sweeps K: including the target's candidates adds $+0.89$ over student-only top-10 at the same K on Qwen3.5-4B, and the Qwen3.5-4B ladder peaks at $K{=}50$. The preferred K shifts across datasets while the tested union-support settings remain close to full-vocabulary distillation; we use the full-vocabulary objective as the default and retain top-K union support as the memory-efficient option, since it avoids materializing full-vocabulary teacher distributions.

B.3 Effect of Original-Image Anchor

To isolate the role of the original-image anchor, we introduce an anchor coefficient λ into the target score:

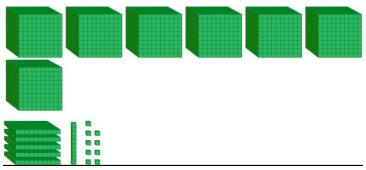
$$a_t^{(\lambda)}(v) = \lambda \log p^J_{\phi,t}(v) + \alpha \widetilde{\Delta}_t(v) = (\lambda + \alpha) \log p^J_{\phi,t}(v) - \alpha \log p^0_{\phi,t}(v), \quad v \in \mathcal{S}_t(\beta). \tag{10}$$

The full method uses $\lambda = 1$, recovering Eq. 8, whereas the no-anchor variant sets $\lambda = 0$ while retaining the same plausibility support. With $\alpha = 1$, its target score becomes $\log p^J_{\phi,t}(v) - \log p^0_{\phi,t}(v)$.

The anchor stabilizes rollout generation while leaving aggregate accuracy unchanged. Table 3 isolates the anchor from contrastive shaping and shows that the aggregate accuracy is comparable across the two configurations. In paired runs differing only in the anchor coefficient, removing the anchor raises the fraction of rollouts that exhaust the generation budget without producing an answer from about one third to about three fifths by the end of training, on both seeds, while validation accuracy and format compliance are unchanged. By keeping the shaped target close to the original prediction, the anchor keeps rollouts terminating and thereby serves as a training-time stability component.

B.4 Qualitative Analysis

MathVista provides a concrete setting for examining whether the aggregate improvement reflects more accurate use of image evidence. **The case study localizes the decisive correction to visual counting.** Figure 6 presents a base-ten block question whose answer depends on the number of



Q: What number is shown?

Ground truth: 7519

Qwen3.5-2B Base predicts **6169** (wrong)
Top row: 6 large cubes → 6,000. Middle: one medium cube, representing 100 → +100. Bottom: 5 flats, 1 rod, 9 units → 69. Total = 6,169.

OPSD predicts **8519** (wrong)
Large cubes (thousands): 7 at the top and 1 below, so 7 + 1 = 8 thousands → 8,000. Then 5 flats (500), 1 rod (10), 9 units (9). Total = 8,519.

VCSD (ours) predicts **7519** (correct)
The top row has 6 large cubes and 1 more below → 7 thousand-cubes (7,000). Plus 5 flats (500), 1 rod (10), and 9 units (9). Total = 7,519.

Figure 6: **Visual counting in a MathVista example with Qwen3.5-2B.** The image and question appear on the left; response excerpts from the base model, OPSD, and VCSD appear on the right. Red highlights counting or place-value errors, and green highlights image-consistent quantities. VCSD correctly identifies the thousand-cubes and combines them with the remaining place-value blocks to obtain the ground-truth answer.

thousand-cubes. The base model misinterprets the place-value blocks and predicts 6169, whereas OPSD overestimates the number of thousand-cubes and predicts 8519. VCSD correctly identifies seven thousand-cubes (six in the top row and one below), producing the correct answer of 7519. Once this image-grounded quantity is recovered, the remaining place-value calculation is direct. This example is consistent with the conditioning contrast shifting the self-distillation target toward candidates supported by instance-specific visual content.

Figure 7 visualizes the learned image-dependent contrast score $\Delta(v) = \log p(v \mid I) - \log p(v \mid I_{\text{ctrl}})$ for the same generated response. **VCSD concentrates additional contrast on image-grounded evidence.** The base model and OPSD exhibit highly similar contrast patterns, whereas VCSD assigns stronger positive contrast to visual concepts such as *roof*, *shingles*, *food*, *beige*, and *sign*, while suppressing less informative tokens. The bottom row isolates the additional contrast learned over OPSD and shows that this change is localized to semantically meaningful visual concepts rather than distributed uniformly across the sequence. This pattern supports the hypothesis that contrastive target shaping reallocates probability mass toward image-dependent tokens instead of merely amplifying existing language preferences.

C Theoretical Perspective

The conditional log-ratio in Equation 5 admits two complementary interpretations. First, it can be viewed as a dense token-level reward measuring the predictive support attributable to instance-specific visual content. Second, it provides a controlled approximation to the conditional pointwise mutual information between a candidate token and the observed image, with the content-erased prediction serving as a surrogate for the prediction without instance-specific visual information.

One-Step KL-Regularized Policy Update. Let the original-image teacher distribution restricted and renormalized over the plausibility support be

$$\bar{p}^{J}_{\phi,t}(v) = \frac{\mathbf{1}[v \in \mathcal{S}_t(\beta)]\, p^{J}_{\phi,t}(v)}{\sum_{u \in \mathcal{S}_t(\beta)} p^{J}_{\phi,t}(u)}. \tag{11}$$

We regard the conditioning contrast

$$r^{\text{vis}}_t(v) := \widetilde{\Delta}_t(v) \tag{12}$$

as an implicit visual-evidence reward.

Remark 1. For each student-generated prefix, the contrast-shaped target in Equation 7 is the unique solution to the following optimization problem:

$$q^{\star}_t = \arg\max_{q \in \Pi(\mathcal{S}_t(\beta))} \left\{ \alpha \mathbb{E}_{v \sim q}\left[r^{\text{vis}}_t(v) \right] - D_{\text{KL}}\left(q \,\|\, \bar{p}^{J}_{\phi,t} \right) \right\}. \tag{13}$$



Figure 7: **Token-level image-dependent contrast.** The base model, OPSD, and VCSD score the same response at matching prefixes under the displayed image and a black-image control. Token colors encode the log-probability difference $\Delta(v) = \log p(v \mid I) - \log p(v \mid I_{\text{ctrl}})$: warmer colors indicate greater preference under the original image. The bottom row shows $\Delta_{\text{VCSD}} - \Delta_{\text{OPSD}}$, isolating the change relative to OPSD. Stronger positive changes on visual terms such as *roof*, *shingles*, and *beige* illustrate how the learned contrast varies across tokens.

Here, $\Pi(\mathcal{S}_t(\beta))$ denotes the probability simplex over $\mathcal{S}_t(\beta)$. The closed-form solution is

$$q_t^\star(v) = \frac{\bar{p}_{\phi,t}^J(v) \exp\left(\alpha r_t^{\text{vis}}(v)\right)}{\sum_{u \in \mathcal{S}_t(\beta)} \bar{p}_{\phi,t}^J(u) \exp\left(\alpha r_t^{\text{vis}}(u)\right)}. \tag{14}$$

Because the support-normalization constant in $\bar{p}_{\phi,t}^J$ cancels between the numerator and denominator, Equation 14 is exactly equivalent to the target defined in Equation 7. The original-image prediction therefore acts as the reference policy, while the conditioning log-ratio provides the reward used to improve it. The plausibility restriction further imposes a hard support constraint, preventing tokens with large likelihood ratios but negligible original-image probability from dominating the improved target. A complete derivation is provided in Appendix D.

Here, "one-step" refers to a single closed-form update from the support-normalized original-image teacher distribution $\bar{p}_{\phi,t}^J$ to the locally improved target $q_t^\star$ at a fixed response prefix. It does not refer to a single gradient step of student optimization.

Importantly, this interpretation does not assume that the original-image policy was historically obtained by applying reinforcement learning to the control-image policy. Rather, the conditional log-ratio defines an implicit reward whose KL-regularized optimum coincides with our contrast-shaped target.

Approximation to Conditional Pointwise Mutual Information. Let $H_t = (P, y_{<t})$ denote the textual context at step t. The conditional pointwise mutual information between token v and image J, conditioned on H_t, takes the form

$$\text{PMI}(v; J \mid H_t) = \log \frac{p(v \mid J, H_t)}{p(v \mid H_t)}. \tag{15}$$

Our conditioning contrast instead takes the form

$$\Delta_t(v) = \log \frac{p_\phi(v \mid J, H_t)}{p_\phi(v \mid J_{\mathrm{ctrl}}, H_t)}. \tag{16}$$

When the content-erased prediction $p_\phi(v \mid J_{\mathrm{ctrl}}, H_t)$ approximates the model's prediction without instance-specific visual information, it serves as a controlled surrogate for $p_\phi(v \mid H_t)$. Under this interpretation, $\Delta_t(v)$ is a contrastive approximation to conditional pointwise mutual information: it measures how much the observed image changes the token's log-likelihood beyond the content-independent prediction. Positive values identify tokens whose likelihood increases in the presence of the observed visual content, whereas values near zero identify tokens whose likelihood is largely preserved after that content is removed. The guarded reward $r_t^{\mathrm{vis}}(v) = \widetilde{\Delta}_t(v)$ follows this interpretation for ordinary tokens, while designated termination tokens are excluded from contrastive adjustment for training stability.

This shows why the proposed objective is useful. Maximizing conditional mutual information favors predictions that remain informative about the observed image after accounting for the textual context, rather than predictions that can be explained primarily by language priors. Accordingly, upweighting tokens with large positive contrast encourages the student to preserve image-dependent evidence in its next-token distribution, while leaving visually insensitive tokens largely unchanged. Combined with the original-image teacher distribution as the reference policy, this yields a conservative form of visual grounding: the method amplifies visually supported distinctions only among tokens that the teacher already considers plausible, instead of rewarding arbitrary image-sensitive outputs. The resulting target therefore promotes stronger dependence on instance-specific visual evidence while retaining the fluency and semantic knowledge encoded by the teacher.

D Proof of Remark 1

We prove Remark 1 from Section C. Fix a student-generated prefix $y_{<t}$, and abbreviate the plausibility support as $\mathcal{S}_t = \mathcal{S}_t(\beta)$. As in Equation 11, let

$$\bar{p}^J_{\phi,t}(v) = \frac{\mathbf{1}[v \in \mathcal{S}_t]\, p^J_{\phi,t}(v)}{\sum_{u \in \mathcal{S}_t} p^J_{\phi,t}(u)} \tag{17}$$

denote the original-image teacher distribution renormalized over $\mathcal{S}_t$, and, as in Equation 12, let

$$r_t^{\mathrm{vis}}(v) = \widetilde{\Delta}_t(v) \tag{18}$$

denote the implicit visual-evidence reward.

Consider the following optimization over $\Pi(\mathcal{S}_t)$, the probability simplex over $\mathcal{S}_t$:

$$\max_{q \in \Pi(\mathcal{S}_t)} \mathcal{J}_t(q) := \alpha \sum_{v \in \mathcal{S}_t} q(v) r_t^{\mathrm{vis}}(v) - \sum_{v \in \mathcal{S}_t} q(v) \log \frac{q(v)}{\bar{p}^J_{\phi,t}(v)}. \tag{19}$$

Existence and uniqueness of the maximizer. Since the teacher distribution is a (temperature-scaled) softmax output, $p^J_{\phi,t}(v) > 0$ for every $v \in \mathcal{V}$; hence, for any $\beta \in [0, 1]$, $\bar{p}^J_{\phi,t}(v) > 0$ for every $v \in \mathcal{S}_t$. Under the convention $0 \log 0 = 0$, the objective $\mathcal{J}_t$ is therefore finite and continuous on $\Pi(\mathcal{S}_t)$; since $\Pi(\mathcal{S}_t)$ is compact, a maximizer exists. For uniqueness, decompose the objective as

$$\mathcal{J}_t(q) = \sum_{v \in \mathcal{S}_t} q(v) \left[\alpha r_t^{\mathrm{vis}}(v) + \log \bar{p}^J_{\phi,t}(v)\right] - \sum_{v \in \mathcal{S}_t} q(v) \log q(v). \tag{20}$$

The first sum is linear in q, with finite coefficients because $\bar{p}^J_{\phi,t}$ is strictly positive on $\mathcal{S}_t$, while the remaining term $-\sum_{v \in \mathcal{S}_t} q(v) \log q(v)$ is the Shannon entropy of q, which is strictly concave on $\Pi(\mathcal{S}_t)$. Hence $\mathcal{J}_t$ is strictly concave, and its maximizer is unique.

Deriving the candidate solution. We first solve the problem subject only to the normalization constraint $\sum_{v\in\mathcal{S}_t} q(v) = 1$, and verify afterwards that the omitted nonnegativity constraints $q(v) \geq 0$ are inactive at the resulting solution. Introduce a Lagrange multiplier λ for the normalization constraint:

$$\mathcal{F}(q,\lambda) = \alpha \sum_{v\in\mathcal{S}_t} q(v) r_t^{\mathrm{vis}}(v) - \sum_{v\in\mathcal{S}_t} q(v) \log \frac{q(v)}{\bar{p}^J_{\phi,t}(v)} + \lambda \left(\sum_{v\in\mathcal{S}_t} q(v) - 1 \right). \tag{21}$$

For every $v \in \mathcal{S}_t$, the stationarity condition is

$$\frac{\partial \mathcal{F}}{\partial q(v)} = \alpha r_t^{\mathrm{vis}}(v) - \log \frac{q(v)}{\bar{p}^J_{\phi,t}(v)} - 1 + \lambda = 0. \tag{22}$$

Rearranging gives

$$\log \frac{q(v)}{\bar{p}^J_{\phi,t}(v)} = \alpha r_t^{\mathrm{vis}}(v) + \lambda - 1, \tag{23}$$

and hence

$$q(v) = \bar{p}^J_{\phi,t}(v) \exp\left(\alpha r_t^{\mathrm{vis}}(v)\right) \exp(\lambda - 1). \tag{24}$$

The factor $\exp(\lambda - 1)$ is shared across all candidates and is determined by normalization. Therefore,

$$q_t^\star(v) = \frac{\bar{p}^J_{\phi,t}(v) \exp\left(\alpha r_t^{\mathrm{vis}}(v)\right)}{\sum_{u\in\mathcal{S}_t} \bar{p}^J_{\phi,t}(u) \exp\left(\alpha r_t^{\mathrm{vis}}(u)\right)}. \tag{25}$$

Verifying global optimality. Every factor on the right-hand side of Equation 25 is strictly positive, so $q_t^\star(v) > 0$ for all $v \in \mathcal{S}_t$, and the nonnegativity constraints omitted above are strictly satisfied and hence inactive at $q_t^\star$. Together with the stationarity condition in Equation 22 and primal feasibility, $q_t^\star$ therefore satisfies the Karush–Kuhn–Tucker (KKT) conditions of the problem in Equation 19, with zero multipliers on the inactive nonnegativity constraints. Since the objective is concave and all constraints are affine, the KKT conditions are sufficient for global optimality, so $q_t^\star$ is the maximizer of Equation 19; by the strict concavity established above, it is the unique one. Intuitively, no maximizer can place zero mass on any candidate in $\mathcal{S}_t$, since $\partial \mathcal{J}_t / \partial q(v) \to +\infty$ as $q(v) \to 0^+$.

Equivalence with the contrast-shaped target. For $v \in \mathcal{S}_t$, substituting

$$\bar{p}^J_{\phi,t}(v) = \frac{p^J_{\phi,t}(v)}{\sum_{u\in\mathcal{S}_t} p^J_{\phi,t}(u)} \tag{26}$$

into Equation 25, the support-normalization constant appears in both the numerator and denominator and therefore cancels. We obtain

$$q_t^\star(v) = \frac{\mathbf{1}[v \in \mathcal{S}_t]\, p^J_{\phi,t}(v) \exp\left(\alpha \widetilde{\Delta}_t(v)\right)}{\sum_{u\in\mathcal{S}_t} p^J_{\phi,t}(u) \exp\left(\alpha \widetilde{\Delta}_t(u)\right)}. \tag{27}$$

This is exactly the contrast-shaped target in Equation 7, which proves the result.

For $\alpha > 0$, an equivalent form of the objective is

$$q_t^\star = \arg\max_{q\in\Pi(\mathcal{S}_t)} \left\{ \mathbb{E}_{v\sim q}\left[r_t^{\mathrm{vis}}(v)\right] - \frac{1}{\alpha} D_{\mathrm{KL}}\left(q \,\|\, \bar{p}^J_{\phi,t}\right) \right\}. \tag{28}$$

Thus, increasing α assigns greater weight to the visual-evidence reward relative to the KL regularizer, while the limit $\alpha \to 0$ recovers the support-normalized original-image teacher distribution.

E AI USE STATEMENT

We used generative AI tools to improve the manuscript's readability and linguistic clarity, assist with drafting and revising the discussion of experimental results, review the implementation of our method, and revise figure-generation code.

The authors reviewed the AI-assisted text and code-review suggestions. We checked that manuscript revisions were consistent with our intended meaning and reported results, and assessed code suggestions against the corresponding implementation. We take full responsibility for all content of this work, including its claims, theoretical arguments, experimental analysis, code, and figures.