LEGO-OPD: FACTORIZED TEACHER COMPOSITION FOR MULTIMODAL ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal on-policy distillation (OPD) aims to improve visual grounding while preserving the strong reasoning capabilities of language models. Recent multi-teacher approaches combine LLM and VLM teachers to provide complementary supervision. However, directly using a VLM's full predictive distribution entangles its visual grounding signal with its own language prior, preventing the grounding information from being transferred independently. Conversely, increasing the strength of visual supervision can improve perception but may overemphasize visual evidence and degrade language reasoning. To address this trade-off, we introduce **LEGO-OPD**, which selectively composes factors from a **L**anguage **E**xpert and a **G**rounding expert into **O**ne teacher distribution for multimodal OPD. Under a generalized Bayesian formulation, the language expert provides a prior over candidate tokens, while the grounding expert contributes a visual likelihood that updates this prior, rather than transferring its complete predictive distribution. This factorized composition allows language reasoning and visual grounding to be controlled independently. We further introduce adaptive calibration to determine how strongly the visual likelihood should update the language prior at each decoding prefix. Specifically, LEGO-OPD uses the grounding expert's image-induced prediction shift as a prefix-dependent reference, preventing both insufficient and excessive visual supervision. Experiments with Qwen3 models show that LEGO-OPD consistently outperforms the evaluated single- and multi-teacher OPD baselines on both multimodal and text-only reasoning tasks. Moreover, it improves the initial student's visual perception while preserving text-only reasoning.

1 INTRODUCTION

On-policy distillation (OPD) is an effective post-training framework for multimodal reasoning in vision-language models (VLMs), distilling teacher predictions at student-generated prefixes (Agarwal et al., 2024; Yin et al., 2026). Effective multimodal OPD requires reliable visual grounding and strong reasoning over grounded information.

Recent approaches largely follow two directions. The first introduces additional specialized teachers to provide complementary supervision, for example by combining external large language models (LLMs) for reasoning with VLM teachers for multimodal guidance (Kwai Keye Team et al., 2026; Yin et al., 2026). However, directly combining their complete predictive distributions entangles each teacher's language prior and grounding signal, making it difficult to selectively transfer their complementary strengths. The second direction strengthens the visual component of the distillation signal to improve grounding (Liu et al., 2026; Liang et al., 2026; Yoon et al., 2026). While stronger visual supervision can improve perception, excessive emphasis on visual evidence may distort the underlying language prior and degrade reasoning performance. Together, these limitations highlight the need for a distillation framework that can separately control language reasoning and visual grounding while combining them into a coherent teacher distribution.

In this paper, we introduce LEGO-OPD, which constructs a single teacher distribution by composing complementary factors from independently selected experts. Our method consists of two components. First, we factorize the teacher into a language prior and a visual grounding factor. An external LLM serves as the language expert and provides a prior distribution over candidate tokens given the textual context, while a VLM serves as the grounding expert and evaluates their compatibility with

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.

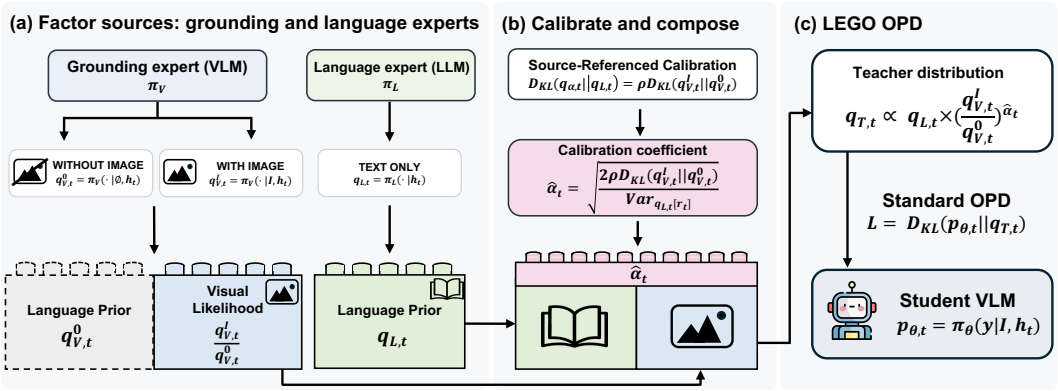


Figure 1: **Overview of LEGO-OPD.** (a) A grounding expert (VLM) supplies a visual likelihood ratio, while a language expert (LLM) supplies a language prior. (b) Source-referenced calibration adjusts the contribution of the visual likelihood ratio. (c) The calibrated factors are composed into a single teacher distribution for on-policy distillation of the student VLM.

the image. This composition admits a generalized Bayesian interpretation in which the language distribution acts as a prior and the visual signal acts as a likelihood that updates the prior into a unified teacher distribution. Second, we introduce source-referenced calibration to determine how strongly the visual factor should modify the language prior at each prefix. Existing prediction-space composition methods, such as DExperts, proxy-tuning, and visual contrastive decoding, typically rely on fixed composition coefficients (Liu et al., 2021; 2024; Leng et al., 2024). However, a fixed coefficient can induce substantially different prediction shifts depending on the underlying prior and decoding prefix. We therefore use the grounding expert's *source KL*, which measures the change in its prediction induced by the image, as a prefix-dependent reference for calibrating the visual update. The scaling coefficient is adapted so that the magnitude of the visual update is normalized relative to this reference, thereby balancing visual grounding against language-prior supervision. A global multiplier on the reference further controls the desired degree of visual emphasis. The composed teacher supervises the student through standard reverse-KL on-policy distillation.

Experiments with Qwen3 models show that LEGO-OPD consistently outperforms the evaluated OPD baselines on both multimodal and text-only reasoning tasks. Further analysis shows that adaptive calibration strengthens visual grounding while preserving text-only reasoning ability. Ablation studies confirm that both factorized teacher composition and adaptive calibration are essential to achieving these improvements.

Our contributions are summarized as follows.

- **Factorized teacher composition.** We compose a language prior and a visual likelihood estimate from independently selected experts into a single OPD teacher through generalized Bayesian updating.
- **Source-referenced calibration.** We adaptively determine the visual update's strength at each prefix, with a global scale controlling the overall visual emphasis.
- **Empirical evaluation.** We demonstrate improvements over the evaluated OPD baselines in both multimodal and text-only reasoning, and show that adaptive calibration enables stronger visual perception without sacrificing text-only reasoning performance.

2 Related Works and Background

2.1 Related Work

Heterogeneous Teachers for Multimodal Distillation. Recent OPD methods use specialized teachers with complementary strengths, an approach also explored in multimodal distillation. These methods differ in the granularity of teacher selection and aggregation. For example, M-OPD routes examples to domain-specific teachers (Ma et al., 2026), and Keye-VL extends routing across modalities and tasks (Kwai Keye Team et al., 2026). ViGOS uses differently conditioned self-teachers for

perception and reasoning segments (Wang et al., 2026), while H-OPD combines LLM and VLM predictions by confidence at each token (Yin et al., 2026). However, external LLM supervision can require additional text-only training data or image descriptions (Appendix C.5). Routing teachers or mixing their complete predictions still entangles the multimodal teacher's language prior and visual grounding signal, limiting independent transfer of each expert's strengths.

Visual Supervision in Multimodal Distillation. A complementary line of multimodal OPD work strengthens the visual information carried by the supervision signal. VA-OPD uses sensitivity to visual detail to reweight rollouts and balance supervision across token groups (Liu et al., 2026). VCSD refines teacher targets through original-image and content-erased contrasts, but reports performance degradation under excessive visual emphasis (Liang et al., 2026). Decomposed OPD combines an auxiliary visual objective with a language preservation loss to mitigate interference from excessive visual emphasis (Yoon et al., 2026). These findings highlight the importance of balancing visual and language-prior supervision.

Composition in Prediction Space. Prediction-space composition offers a way to transfer signals across models. DExperts uses expert and anti-expert predictions (Liu et al., 2021), proxy-tuning transfers prediction shifts from smaller tuned models (Liu et al., 2024), and UniR adds learned reasoning-module logits to a frozen LLM (Kim et al., 2026). Visual contrastive decoding similarly amplifies prediction differences across image conditions (Leng et al., 2024). However, these methods use fixed composition coefficients, leaving the choice of suitable composition strengths to predefined settings or empirical tuning.

LEGO-OPD composes complementary expert factors instead of complete predictions. Source-referenced calibration selects composition strength by adapting the visual update at each prefix.

2.2 Generalized Bayesian Updating

Generalized Bayesian updating incorporates new information into a prior without requiring a complete likelihood model for the observations. The framework of Bissiri et al. (2016) represents this information through an observation-dependent loss L. We use the equivalent evidence-score convention $s(v) := -L(v)$, where higher scores indicate greater support from the observed information.

For a strictly positive prior p_0 over a finite candidate space $\mathcal{V}$, a finite evidence score s, and an evidence scale $\alpha \geq 0$, the update is

$$q^\star_\alpha = \arg\max_{q \in \Delta_\mathcal{V}} \left\{ \alpha \mathbb{E}_{v \sim q}[s(v)] - D_{\mathrm{KL}}(q \| p_0) \right\}, \tag{1}$$

where $\Delta_\mathcal{V}$ is the set of probability distributions over $\mathcal{V}$. Its unique solution is

$$q^\star_\alpha(v) = \frac{p_0(v) \exp\{\alpha s(v)\}}{\sum_{u \in \mathcal{V}} p_0(u) \exp\{\alpha s(u)\}}. \tag{2}$$

The coefficient α specifies the influence of the evidence relative to the prior. When $s(v)$ is the log-likelihood of the observations under a specified probabilistic model and $\alpha = 1$, the update recovers that model's Bayesian posterior. More generally, s can encode information without specifying a complete sampling distribution.

2.3 On-Policy Distillation

On-policy distillation supervises a student policy π_θ at student-generated prefixes (Agarwal et al., 2024). For $(x, I) \sim \mathcal{D}$, let h_t contain the system prompt, question, and response prefix $y_{<t}$. Write $p_{\theta,t}(v) := \pi_\theta(v \mid I, h_t)$ and $q_t(v)$ for the student and teacher distributions over $\mathcal{V}$ at this prefix. At update k, reverse-KL OPD minimizes

$$\mathcal{L}_k(\theta) = \mathbb{E}_{\substack{(x,I)\sim\mathcal{D} \\ y \sim \pi_{\theta_k}(\cdot|x,I)}} \left[\frac{1}{|y|} \sum_{t=1}^{|y|} D_{\mathrm{KL}}(p_{\theta,t} \,\|\, q_t) \right], \tag{3}$$

where $|y|$ is the response length. Prefixes and teacher targets are fixed during each update.

3 METHOD

3.1 LEGO-OPD: MULTIMODAL OPD WITH FACTORIZED TEACHER COMPOSITION

Notation. Let π_L and π_V denote the frozen policies of an independently selected language expert (LLM) and grounding expert (VLM), respectively. For a question–image pair (x, I), let h_t contain the system prompt, question, and student-generated prefix $y_{<t}$. Both experts share the token vocabulary $\mathcal{V}$ and are evaluated at the same textual context. For $v \in \mathcal{V}$, define $q_{L,t}(v) := \pi_L(v \mid h_t)$, $q^I_{V,t}(v) := \pi_V(v \mid I, h_t)$, and $q^0_{V,t}(v) := \pi_V(v \mid \emptyset, h_t)$, where $\emptyset$ denotes the no-image condition. All three distributions are obtained by applying softmax to finite logits over the finite vocabulary $\mathcal{V}$.

Overall Framework. Figure 1 illustrates our LEGO-OPD for on-policy VLM distillation using factorized teacher composition. A grounding expert (VLM) supplies a visual likelihood ratio, while a language expert (LLM) supplies a language prior. Additionally, source-referenced calibration adjusts the contribution of the visual likelihood ratio, while the calibrated factors are composed into a single teacher distribution for on-policy distillation of the student VLM.

Specifically, let $z_{L,t}$, $z^I_{V,t}$, and $z^0_{V,t}$ denote logits whose softmaxes give $q_{L,t}$, $q^I_{V,t}$, and $q^0_{V,t}$, respectively. Using the adaptive coefficient from Eq. (13), which will be explained soon, the composition yields the final teacher distribution

$$q_{T,t} := q_{\widehat{\alpha}_t,t} = \mathrm{softmax}_{\mathcal{V}}\left(z_{L,t} + \widehat{\alpha}_t\left[z^I_{V,t} - z^0_{V,t}\right]\right). \tag{4}$$

Branch-specific log-normalizers cancel in the final softmax, making the logit- and probability-space compositions equivalent. For numerical stability, the implementation uses a stabilized residual and regularized coefficient estimates, with statistics computed over high-probability tokens to limit sensitivity to tail predictions. Details are in Appendix B.

Finally, let π_θ denote the multimodal student policy, with next-token distribution $p_{\theta,t}(v) := \pi_\theta(v \mid I, h_t)$ at each student-generated prefix. The language expert π_L and grounding expert π_V provide the factors used to construct $q_{T,t}$ in Eq. (4). We use this composed teacher distribution as the target for the reverse-KL OPD objective in Eq. (3), holding the sampled prefixes and teacher targets fixed during each student update.

3.2 LEGO-OPD: THEORETICAL BACKGROUND

In this section, we provide theoretical background how we come up with the factored teacher Eq. (4) as our OPD target in our LEGO-OPD.

Bayesian Factorization and Cross-Expert Sourcing. For a joint distribution over the next token and image, Bayes' rule decomposes a multimodal prediction into two complementary factors,

$$P(v \mid I, h_t) \propto \underbrace{P(v \mid h_t)}_{\text{language prior}}\,\underbrace{P(I \mid v, h_t)}_{\text{visual likelihood}}. \tag{5}$$

The language prior supports reasoning based on the textual context. The visual likelihood supports grounding by reweighting tokens based on their compatibility with the image. We source the two factors from independently selected experts to draw on their complementary strengths in reasoning and grounding. The language expert π_L provides the language prior $q_{L,t}$ as its next-token distribution conditioned on the textual context h_t.

Although the autoregressive grounding expert π_V cannot directly evaluate $P(I \mid v, h_t)$, Bayes' rule (Eq. (5)) gives

$$P(I \mid v, h_t) \propto \frac{P(v \mid I, h_t)}{P(v \mid h_t)}. \tag{6}$$

To estimate the visual likelihood factor, we use the ratio of the grounding expert's image-conditioned and image-free predictions, following Decomposed OPD (Yoon et al., 2026). The estimate and its logarithm are defined as follows.

$$\widehat{\ell}_t(v) := \frac{q^I_{V,t}(v)}{q^0_{V,t}(v)}, \qquad r_t(v) := \log \widehat{\ell}_t(v) = \log q^I_{V,t}(v) - \log q^0_{V,t}(v). \tag{7}$$

In particular, $q^0_{V,t}$ is obtained by removing the image input rather than marginalizing over images. The ratio in Eq. (7) therefore need not correspond to the likelihood term of an underlying joint distribution. Instead, it measures the image-induced change in the grounding expert's own predictive distribution and is inherently model-dependent. We thus regard $\widehat{\ell}_t$ as a proxy for visual evidence, which we refer to as the visual likelihood estimate, and its logarithm r_t as the visual residual.

Generalized Bayesian Teacher Composition. We obtain the language prior $q_{L,t}$ and the visual likelihood estimate $\widehat{\ell}_t$ from independently selected experts. Their normalized product is therefore not guaranteed to represent an exact Bayesian posterior. We therefore adopt generalized Bayesian updating, as described in Section 2.2, to update the language prior using the evidence provided by the visual likelihood estimate.

More specifically, we use the visual residual $r_t(v) = \log \widehat{\ell}_t(v)$ as the evidence score. Higher values indicate greater relative support for a candidate token in the presence of the image. At each student-generated prefix, substituting $p_0 = q_{L,t}$ and $s = r_t$ into Eq. (1) defines the teacher distribution for an evidence scale $\alpha_t \geq 0$,

$$q_{\alpha,t} := \arg\max_{q \in \Delta_{\mathcal{V}}} \left\{ \alpha_t \mathbb{E}_{v \sim q}[r_t(v)] - D_{\mathrm{KL}}(q \| q_{L,t}) \right\}. \tag{8}$$

The objective balances the expected evidence score against KL deviation from the language prior. The unique solution for Eq. (8) is

$$q_{\alpha,t}(v) = \frac{q_{L,t}(v) \exp\{\alpha_t r_t(v)\}}{Z_t(\alpha)}, \qquad Z_t(\alpha) := \sum_{u \in \mathcal{V}} q_{L,t}(u) \exp\{\alpha_t r_t(u)\}, \tag{9}$$

where $Z_t(\alpha)$ is the normalizing constant. By plugging in Eq. (7) into Eq. (9), our teacher assumes the specific form:

$$q_{\alpha,t}(v) = \frac{1}{Z_t(\alpha)} q_{L,t}(v) \left(\frac{q^I_{V,t}(v)}{q^0_{V,t}(v)} \right)^{\alpha_t} \tag{10}$$

whose logit expression leads to the final form Eq. (4) of our factored composite teacher. Accordingly, the generalized Bayesian posterior defines the teacher distribution by combining the language prior from the language expert with the visual likelihood from the grounding expert. Any candidate-independent multiplicative constant in the visual likelihood cancels after normalization.

The evidence scale α_t determines the strength of the visual update relative to the language prior. We next describe how α_t is adaptively selected at each student-generated prefix.

Source-Referenced Calibration. We interpret α_t as a calibration coefficient for composing the language prior and visual likelihood estimate. Motivated by the limitations of a fixed coefficient illustrated in Appendix D.1, we use the grounding expert's image-induced prediction shift as a reference for selecting α_t at each prefix. This shift is measured by the *source KL*, $D_{\mathrm{KL}}(q^I_{V,t} \| q^0_{V,t})$. More specifically, at each prefix, we choose α_t so that the composed teacher's KL divergence from the language prior matches the scaled source KL:

$$D_{\mathrm{KL}}(q_{\alpha_t,t} \| q_{L,t}) = \rho \, D_{\mathrm{KL}}(q^I_{V,t} \| q^0_{V,t}), \tag{11}$$

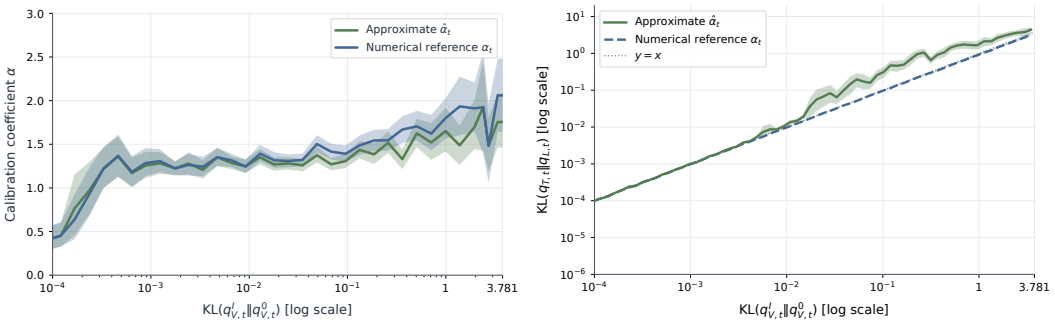
where $\rho \geq 0$ is the KL-budget scale shared across prefixes. For a nonconstant residual, $D_{\mathrm{KL}}(q_{\alpha,t} \| q_{L,t})$ is continuous and strictly increasing in $\alpha > 0$, so Eq. (11) admits a unique finite positive solution whenever the requested budget lies within its attainable range (Appendix A.2).

Solving for this coefficient exactly would require a one-dimensional root search at every generated position. We therefore locally approximate $D_{\mathrm{KL}}(q_{\alpha,t} \| q_{L,t})$ around $\alpha = 0$, where the composed teacher equals the language prior. A second-order Taylor expansion gives

$$D_{\mathrm{KL}}(q_{\alpha,t} \| q_{L,t}) = \tfrac{1}{2} \mathrm{Var}_{q_{L,t}}[r_t] \alpha^2 + O(\alpha^3). \tag{12}$$

Matching the quadratic term to the KL budget gives the approximate coefficient

$$\widehat{\alpha}_t := \sqrt{\frac{2\rho D_{\mathrm{KL}}(q^I_{V,t} \| q^0_{V,t})}{\mathrm{Var}_{q_{L,t}}[r_t]}}. \tag{13}$$



(a) Calibration coefficient (b) Induced teacher KL

Figure 2: **Local approximation error in expert composition.** Curves show means within source-KL bins at $\rho = 1$. The dotted $y = x$ line denotes source-KL matching. Shaded bands indicate question-bootstrap 95% confidence intervals.

For a fixed KL budget, this approximation assigns a smaller coefficient when the residual varies more strongly under the receiving language prior. We set $\widehat{\alpha}_t = 0$ when the variance is zero. Because the approximation is local, the achieved KL may differ from the requested budget for finite updates. The derivation and approximation properties are detailed in Appendix A.3.

To assess the local approximation, we compute calibration coefficients using the approximation and a numerical KL-matching solver on shared prefixes for the 4B language and grounding experts. We use 24,576 shared prefixes from 768 rollouts on 256 MMMU (Yue et al., 2024) questions, using student checkpoints at steps 0, 100, and 200. Experimental details are provided in Appendix D.5. Figure 2(a) shows that the mean approximate coefficient broadly tracks the numerical solution across source-KL intervals, with larger deviations at larger source KL. Figure 2(b) shows that the mean teacher KL under $\widehat{\alpha}_t$ closely matches the source KL for small updates but exceeds it for larger updates. This is consistent with the local second-order approximation (Equation 12). We further examine the effects of approximation error in a practical distillation setting in Appendix D.6.

Loss Decomposition and its Interpretation. To examine the teacher factors' contributions to the OPD objective, we express the student's prediction in a factorized form analogous to the teacher composition. Let $p^0_{\theta,t}(v) := \pi_\theta(v \mid \emptyset, h_t)$ denote the student's language prior and $r^S_t(v) := \log p_{\theta,t}(v) - \log p^0_{\theta,t}(v)$ its visual residual. Thus, $p_{\theta,t}(v) = p^0_{\theta,t}(v) \exp\{r^S_t(v)\}$. Substituting this factorization and the teacher distribution in Eq. (9) into the reverse-KL objective gives

$$D_{\mathrm{KL}}(p_{\theta,t} \| q_{T,t}) = \underbrace{\mathbb{E}_{p_{\theta,t}}\left[\log \frac{p^0_{\theta,t}(v)}{q_{L,t}(v)}\right]}_{\text{language-prior alignment}} + \underbrace{\mathbb{E}_{p_{\theta,t}}\left[r^S_t(v) - \widehat{\alpha}_t r_t(v)\right]}_{\text{visual-likelihood alignment}} + \log Z_t(\widehat{\alpha}_t). \tag{14}$$

The first and second terms capture the language-prior and visual-residual contributions to student–teacher alignment, respectively, with the grounding expert's residual scaled by $\widehat{\alpha}_t$. The final term normalizes the teacher distribution and is independent of the student parameters at a fixed prefix. Eq. (14) thus expresses student–teacher alignment through language-prior and visual-residual contributions within a single reverse-KL objective.

4 Experiments

4.1 Experimental Setup

Models and training data. For LEGO-OPD, we use Qwen3-VL-2B-Instruct as the student, Qwen3-4B-Instruct-2507 as the language expert, and Qwen3-VL-4B-Instruct as the grounding expert (Yang et al., 2025; Bai et al., 2025). We use the 54,840-example subset of MMFineReason-123K released by H-OPD as the multimodal training data (Lin et al., 2026; Yin et al., 2026). This subset supports comparison with H-OPD and includes the image descriptions required by its language teacher.

Training. All experiments, including ablations, use a common training implementation, student initialization, and optimization settings, except for method-specific components. All student param-

eters are updated for 200 steps using AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-6}, no warmup, and global batch size 128. Prompt and response lengths are limited to 12,384 tokens. Rollouts use temperature 1.0 without top-p or top-k truncation. We follow the top-16 OPD setting of Li et al. (2026). Implementation details are in Appendices B and C.1.

Evaluation. We evaluate multimodal reasoning, text-only reasoning, and visual perception separately, as summarized in Table 1. The multimodal reasoning benchmarks comprise MATH-Vision (Wang et al., 2024a), MathVerse (Zhang et al., 2025), MMMU-Pro (Yue et al., 2025), MathVista (Lu et al., 2024), RBench-M (Guo et al., 2025), and OlympiadBench-MM (He et al., 2024). For text-only evaluation, we use MATH-500 (Lightman et al., 2024), MathVerse-Text Only (Zhang et al., 2025), MMLU-Pro (Wang et al., 2024b), RBench-T (Guo et al., 2025), OlympiadBench-Text (He et al., 2024). To assess visual perception separately from reasoning performance, we evaluate BLINK (Fu et al., 2025) and QBench1,2 (Wu et al., 2024; Zhang et al., 2024). For QBench1,2, we pool the perception-focused Group A questions from both benchmarks (1,495 + 1,000 questions) and report accuracy over the combined set. Except for perception, we average five samples per question. Category averages are unweighted benchmark means. Evaluation details are provided in Appendix C.3.

Baselines. We compare LEGO-OPD with standard OPD, VA-OPD (Liu et al., 2026), M-OPD (Ma et al., 2026), and H-OPD (Yin et al., 2026). For controlled comparison, we match the model scale, optimization settings, and response budget whenever compatible with each method, while preserving its defining supervision structure. M-OPD and H-OPD use both VLM and LLM teachers, whereas standard OPD and VA-OPD rely only on a VLM teacher. Since M-OPD is formulated for domain-routed text distillation, we route our multimodal training data to the VLM teacher and DeepMath to the LLM teacher. VA-OPD retains its multi-response sampling scheme, generating four responses per prompt and therefore observing fewer unique training prompts under the same response budget. Further implementation details are provided in Appendix C.5.

Table 1: Results on multimodal reasoning, perception, and text reasoning. Bold marks the best distilled-student score. Evaluation details and SDs are in Appendices C.3 and E.1.1.

	Teacher	**Initial Student**	**Distillation Baselines**				**Our Method**	
Model / Method	**Qwen3-VL 4B-Instruct**	**Qwen3-VL 2B-Instruct**	**Standard OPD**	**VA-OPD**	**M-OPD**	**H-OPD**	**LEGO-OPD** $\rho = 1$	**LEGO-OPD** $\rho = 2$
Teacher	–	–	Qwen3-VL-4B	Qwen3-VL-4B	Qwen3-VL-4B + Qwen3-4B	Qwen3-VL-4B + Qwen3-4B	Qwen3-VL-4B + Qwen3-4B	Qwen3-VL-4B + Qwen3-4B
Multimodal Reasoning								
MATH-Vision	39.74	27.96	27.43	25.33	24.41	32.37	33.36	**34.28**
MathVerse (Total)	66.70	51.51	52.70	51.66	51.28	55.41	56.96	**57.19**
MathVista	76.26	66.56	68.60	67.36	66.42	67.70	69.62	**70.58**
MMMU-Pro	50.29	37.40	36.98	35.11	35.06	39.36	40.98	**41.38**
RBench-M	32.81	19.61	16.90	16.81	15.67	26.92	29.17	**29.20**
OlympiadBench-MM	63.20	37.60	39.60	37.33	36.53	39.07	44.93	**45.07**
Multimodal Avg.	54.83	40.11	40.37	38.93	38.23	43.47	45.84	**46.28**
Perception								
BLINK	63.49	51.34	52.13	**52.81**	51.87	51.76	52.03	52.55
QBench1,2	76.63	69.18	69.54	**69.70**	68.86	68.98	69.34	69.62
Text Reasoning								
MATH-500	93.04	80.44	78.72	80.20	80.56	81.88	**82.40**	82.36
MathVerse (Text Only)	66.80	57.18	57.44	56.65	56.19	59.44	59.59	**60.23**
MMLU-Pro	62.90	43.59	45.52	45.09	44.98	46.94	47.15	**47.54**
RBench-T	51.92	30.20	28.59	28.76	28.28	37.17	38.59	**40.27**
OlympiadBench-Text	69.41	48.04	48.31	48.99	47.77	52.40	52.31	**53.00**
Text Avg.	68.81	51.89	51.72	51.94	51.56	55.57	56.01	**56.68**

4.2 Main Results

LEGO-OPD at both $\rho = 1$ and $\rho = 2$ outperforms the evaluated OPD baselines in multimodal and text-only reasoning averages (Table 1). Both settings outperform H-OPD on all six multimodal reasoning benchmarks using the same expert pair. Increasing ρ from 1 to 2 further improves both reasoning averages and perception scores. These results support factor composition with adaptive

Table 2: **Adaptive calibration and the KL-budget scale.** Generalized Bayesian composition with fixed coefficients or adaptive calibration. Best scores per group are **green** (fixed) and **blue** (adaptive). Reasoning standard deviations are in Appendix D.2.

	Fixed coefficient			Adaptive		
Benchmark	$\alpha = 1$	$\alpha = 2$	$\alpha = 4$	$\rho = 1$	$\rho = 2$	$\rho = 4$
Multimodal Reasoning						
MATH-Vision	**31.91**	31.25	30.00	33.36	**34.28**	34.01
MathVerse (Total)	53.46	**54.29**	53.41	56.96	57.19	**57.51**
MathVista	64.58	68.98	**69.36**	69.62	**70.58**	70.50
MMMU-Pro	**39.48**	38.87	37.45	40.98	**41.38**	41.36
RBench-M	**26.89**	24.03	17.62	29.17	**29.20**	28.39
OlympiadBench-MM	38.27	**41.87**	41.73	44.93	45.07	**45.33**
Multimodal Avg.	42.43	**43.22**	41.60	45.84	**46.28**	46.18
Perception						
BLINK	51.18	51.29	**51.92**	52.03	**52.55**	52.50
QBench1,2	68.46	68.86	**69.66**	69.34	69.62	**69.70**
Text Reasoning						
MATH-500	**82.24**	82.08	81.52	82.40	82.36	**82.52**
MathVerse (Text Only)	**59.06**	58.63	57.94	59.59	60.23	**60.46**
MMLU-Pro	**47.40**	47.16	47.10	47.15	**47.54**	47.38
RBench-T	**38.21**	35.52	32.85	38.59	**40.27**	38.08
OlympiadBench-Text	**52.20**	51.72	51.19	52.31	**53.00**	51.78
Text Avg.	**55.82**	55.02	54.12	56.01	**56.68**	56.04

calibration as an effective way to transfer the experts' complementary strengths. The performance advantage also holds under greedy decoding, with the highest multimodal and text-only reasoning averages among the evaluated 2B students , as reported in Appendix E.1.2.

Table 2 compares fixed-coefficient composition with adaptive calibration using the same student, expert pair, and training settings. Fixed composition at $\alpha = 2$ approaches H-OPD's performance (Tables 1 and 2). However, the best fixed coefficient varies across benchmarks, leaving its choice to empirical tuning. By contrast, adaptive calibration at $\rho = 1$ outperforms all tested fixed coefficients in both reasoning averages without empirically tuning a fixed composition coefficient. Moreover, the gains are larger in multimodal than in text-only reasoning. This may reflect the need to balance visual grounding with language-based reasoning. Adaptive calibration can strengthen relevant visual supervision while preserving language-prior supervision, benefiting multimodal reasoning.

With fixed coefficients, increasing α improves perception but reduces text-only reasoning performance. Increasing α from 2 to 4 also produces a marked decline in the multimodal reasoning average. This pattern suggests that excessive visual emphasis weakens language-prior supervision. Related effects have been reported as interference between visual and language objectives in Decomposed OPD (Yoon et al., 2026) and performance degradation under excessive contrastive shaping in VCSD (Liang et al., 2026). In contrast, adaptive calibration limits changes to the language prior where the source KL is small. This helps preserve language-prior supervision even under stronger visual emphasis. Across $\rho \in \{1, 2, 4\}$, the larger budgets improve perception relative to $\rho = 1$ while maintaining or improving both reasoning averages. This pattern supports adaptive calibration as a way to increase visual emphasis while preserving reasoning performance. Appendix D.3 additionally describes $\rho = 0$ setting.

4.3 Analysis of Language and Grounding Experts

Table 3 compares the full 4B–4B composition with variants that replace one expert's factor with that of the frozen initial student. To focus on reasoning and grounding, the subset includes multimodal mathematics and challenging reasoning benchmarks alongside their available text-only counterparts. Table 4 reports the full reasoning benchmark results and extends the comparison to additional expert combinations. The experiments use Qwen3-1.7B and Qwen3-4B-Instruct-2507 as language experts and Qwen3-VL-Instruct models of sizes 2B, 4B, and 8B as grounding experts. One expert is varied at a time from the default 4B–4B pairing.

A key observation in Table 3 is that the two experts' benefits combine in the full composition. The language prior supports reasoning across modalities, while the visual residual provides additional

Table 3: **Contributions of the language and grounding experts.** Student (2B) is the frozen initial Qwen3-VL-2B. All runs use $\rho = 1$. Parentheses show pp changes from the full composition.

Teacher composition		Multimodal Reasoning			Text Reasoning	
Language prior	**Visual residual**	**MATH-Vision**	**MathVerse (Total)**	**Olympiad Bench-MM**	**MathVerse (Text Only)**	**Olympiad Bench-Text**
Qwen3-4B	Qwen3-VL-4B	**33.36**	**56.96**	**44.93**	59.59	52.31
Qwen3-4B	Student (2B)	33.03 (−0.33)	54.17 (−2.79)	43.07 (−1.86)	**59.92** (+0.33)	**52.67** (+0.36)
Student (2B)	Qwen3-VL-4B	26.58 (−6.78)	52.69 (−4.27)	37.20 (−7.73)	56.98 (−2.61)	48.07 (−4.24)

Table 4: **Effect of expert choice on student reasoning.** One expert is varied from the shaded Base. All runs use the same 2B student, 200 steps, and $\rho = 1$. Student (2B) denotes the frozen initial student. When used as the language expert, it receives no image. Parentheses show average changes from Base in pp (blue increases, red decreases).

Teacher composition	**Language expert swap**		**Base**	**Grounding expert swap**	
Language expert	LLM 1.7B	Student (2B)	**LLM 4B**	LLM 4B	LLM 4B
Grounding expert	VLM 4B	VLM 4B	**VLM 4B**	Student (2B)	VLM 8B
Multimodal Reasoning					
MATH-Vision	28.36	26.58	33.36	33.03	34.34
MathVerse (Total)	43.89	52.69	56.96	54.17	57.81
MathVista	64.96	68.20	69.62	69.42	70.24
MMMU-Pro	35.11	36.36	40.98	41.93	41.70
RBench-M	22.71	17.53	29.17	28.48	28.90
OlympiadBench-MM	36.00	37.20	44.93	43.07	45.20
Multimodal Avg.	38.51 (−7.33)	39.76 (−6.08)	45.84	45.02 (−0.82)	46.37 (+0.53)
Text Reasoning					
MATH-500	81.52	79.60	82.40	82.36	84.08
MathVerse (Text Only)	55.03	56.98	59.59	59.92	59.59
MMLU-Pro	43.14	44.91	47.15	47.51	46.92
RBench-T	33.05	29.87	38.59	38.46	37.61
OlympiadBench-Text	47.57	48.07	52.31	52.67	53.06
Text Avg.	52.06 (−3.95)	51.89 (−4.12)	56.01	56.18 (+0.17)	56.25 (+0.24)

multimodal gains with little change in text-only performance. This pattern supports the joint transfer of language reasoning and visual grounding to the student.

The expanded results in Table 4 clarify the effects of changing each expert. With the 4B grounding expert fixed, replacing the base 4B LLM with the 1.7B LLM lowers both text-only and multimodal reasoning scores. Conversely, with the 4B language expert fixed, increasing the grounding model from 2B to 4B to 8B improves the multimodal reasoning average, while text-only averages remain nearly unchanged. Settings and standard deviations are in Appendix E.2.

5 Conclusion

We presented **LEGO-OPD**, which composes a language prior and a visual grounding signal into a unified teacher for multimodal on-policy distillation. A generalized Bayesian formulation enables factorized teacher composition, while source-referenced calibration adaptively controls the visual contribution at each student-generated prefix. A local KL approximation yields a closed-form calibration coefficient, with its approximation error assessed against a numerical KL-matching reference. Across Qwen3 models, LEGO-OPD achieves the best average performance among the evaluated single- and multi-teacher OPD baselines on both multimodal and text-only reasoning. Ablations show that the language prior and visual residual provide complementary benefits, and that adaptive calibration outperforms fixed weighting in both reasoning averages while allowing stronger visual emphasis without degrading reasoning. Overall, LEGO-OPD provides an effective framework for balancing language reasoning and visual grounding in multimodal distillation. Limitations are discussed in Appendix F.

REPRODUCIBILITY STATEMENT

Section 4.1 describes the models, training data, optimization settings, and evaluation benchmarks used in our experiments. Appendix A provides the theoretical analysis and proofs, and Appendix B specifies the practical teacher construction and student update of LEGO-OPD. Appendix C details the training settings, evaluation protocols, prompt templates, and baseline implementations, and Appendix D gives the configurations for the calibration ablations and diagnostics. The code is available at `https://anonymous.4open.science/r/LEGO_OPD-51C1`.

AI USE STATEMENT

We used generative AI tools as interactive research assistants throughout this work. Our use was mainly iterative. We used them to explore alternative formulations, examine our reasoning, identify ambiguities or potential weaknesses, and obtain feedback as the work developed. They also helped broaden our literature review, refine translation and English scientific writing, and accelerate implementation and debugging.

Generative AI tools also supported discussion of theoretical formulations and mathematical derivations, experimental design, result interpretation, and code development. These tools informed the research process, while the authors set the research direction, made the methodological and experimental decisions, and determined the final interpretations. AI-assisted suggestions were treated as material for further examination rather than authoritative outputs. Mathematical arguments were checked by the authors, AI-assisted implementations were manually reviewed and tested against the intended methods and experimental settings, and relevant claims and citations were verified against their original sources. The authors repeatedly revised the manuscript using AI tools for discussion, critique, translation, and editing while continuing to develop and examine the underlying arguments themselves. Generative AI was also used as a documented part of the evaluation pipeline for semantic answer interpretation, as described in Appendix C.3.

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 B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024. URL `https://proceedings.iclr.cc/paper_files/paper/2024/file/5be69a584901a26c521c2b51e40a4c20-Paper-Conference.pdf`.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report, 2025. URL `https://arxiv.org/abs/2511.21631`.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130, 11 2016. ISSN 1369-7412. doi: 10.1111/rssb.12158. URL `https://doi.org/10.1111/rssb.12158`.

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 Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, volume 15081 of *Lecture Notes*

in Computer Science, pp. 148–166, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73337-6. doi: 10.1007/978-3-031-73337-6_9. URL https://doi.org/10.1007/978-3-031-73337-6_9.

Meng-Hao Guo, Jiajun Xu, Yi Zhang, Jiaxi Song, Haoyang Peng, Yi-Xuan Deng, Xinzhi Dong, Kiyohiro Nakayama, Zhengyang Geng, Chen Wang, Bolin Ni, Guo-Wei Yang, Yongming Rao, Houwen Peng, Han Hu, Gordon Wetzstein, and Shi-Min Hu. RBench: Graduate-level multi-disciplinary benchmarks for LLM & MLLM complex reasoning evaluation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21184–21201. PMLR, 2025. URL https://proceedings.mlr.press/v267/guo25r.html.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. DeepMath-103K: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kHB5Te5IWm.

Jaemin Kim, Hangeol Chang, Hyunmin Hwang, Choonghan Kim, and Jong Chul Ye. Universal reasoner: A single, composable plug-and-play reasoner for frozen LLMs. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=JheeAwquCb.

Kwai Keye Team, Bin Wen, Changyi Liu, Chengru Song, Chongling Rao, Guowang Zhang, Han Li, Haonan Fan, Hengrui Ju, Jiankang Chen, Jiapeng Chen, Jiawei Yuan, Kaixuan Yang, Kaiyu Jiang, Kun Gai, Lingzhi Zhou, Na Nie, Sen Na, Tianke Zhang, Tingting Gao, Xuanyu Zheng, Yulong Chen, Fan Yang, Haixuan Gao, Lele Yang, Mingqiao Liu, Muxi Diao, Qi Zhang, Qile Su, Wei Chen, Wentao Hong, Xingyu Lu, Yancheng Long, Yankai Yang, Yingxin Li, Yiyang Fan, Yu Xia, Yuzhe Chen, Ziliang Lai, Chuan Yi, Haonan Jia, Tianming Liang, Weixin Xu, Xiaoxiao Ma, Yang Tian, Yufei Han, Feng Han, Hang Li, Jing Wang, Jinghui Jia, Junmin Chen, Junyu Shi, and Ruilin Zhang. Kwai Keye-VL-2.0 technical report, 2026. URL https://arxiv.org/abs/2606.10651.

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 (CVPR)*, pp. 13872–13882, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan ang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe, 2026. URL https://arxiv.org/abs/2604.13016.

Yijun Liang, Yunjie Tian, Yijiang Li, Yuqi Jia, Furong Huang, Tianyi Zhou, and Di Fu. Visual contrastive self-distillation, 2026. URL https://arxiv.org/abs/2607.21556.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.),

International Conference on Learning Representations, volume 2024, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.

Honglin Lin, Zheng Liu, Yun Zhu, Chonghan Qin, Juekai Lin, Xiaoran Shang, Conghui He, Wentao Zhang, and Lijun Wu. MMFineReason: Closing the multimodal reasoning gap via open data-centric methods, 2026. URL https://arxiv.org/abs/2601.21821.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6691–6706, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.522. URL https://aclanthology.org/2021.acl-long.522/.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. Tuning language models by proxy. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=dribhnhm1i.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, and Shu Wu. Visual-advantage on-policy distillation for vision-language models, 2026. URL https://arxiv.org/abs/2605.21924.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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 B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/663bce02a0050c4a11f1eb8a7f1429d3-Paper-Conference.pdf.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. MOPD: Multi-teacher on-policy distillation for capability integration in LLM post-training, 2026. URL https://arxiv.org/abs/2606.30406.

Qwen Team. Qwen3-4B-Instruct-2507: Model card. Hugging Face model card, 2025. URL https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507.

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 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95095–95169. Curran Associates, Inc., 2024a. doi: 10.52202/079017-3014. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad0edc7d5fa1a783f063646968b7315b-Paper-Datasets_and_Benchmarks_Track.pdf.

Sihan Wang, Xiyao Liu, Lianqing Liu, and Zhi Han. Seeing before reasoning: Decoupling perception and reasoning for shortcut-resilient multimodal on-policy self-distillation, 2026. URL https://arxiv.org/abs/2606.19120.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290. Curran Associates, Inc., 2024b. doi: 10.52202/079017-3018.

URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad236edc564f3e3156e1b2feafb99a24-Paper-Datasets_and_Benchmarks_Track.pdf.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-Bench: A benchmark for general-purpose foundation models on low-level vision. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 12547–12573, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/363d4c97bb411e4b07612915b76c06ae-Paper-Conference.pdf.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Qixiang Yin, Huanjin Yao, Yuchen Cai, Jianghao Chen, Ziyi Wang, Min Yang, Fei Su, and Zhicheng Zhao. H-OPD: Confidence aware heterogeneous multi-teacher multimodal on-policy distillation, 2026. URL https://arxiv.org/abs/2607.02592.

Hee Suk Yoon, Eunseop Yoon, Jaehyun Jang, SooHwan Eom, Ji Woo Hong, Mark A. Hasegawa-Johnson, Qi Dai, Chong Luo, and Chang D. Yoo. Decomposed on-policy distillation for vision-language reasoning: Steering gradients for visual grounding. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Jb1YBPqyb7.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.736. URL https://aclanthology.org/2025.acl-long.736/.

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 Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, volume 15066 of *Lecture Notes in Computer Science*, pp. 169–186, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73242-3. doi: 10.1007/978-3-031-73242-3_10. URL https://doi.org/10.1007/978-3-031-73242-3_10.

Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. Q-Bench+: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *IEEE Trans-*

actions on Pattern Analysis and Machine Intelligence, 46(12):10404–10418, 2024. doi: 10.1109/TPAMI.2024.3445770. URL https://doi.org/10.1109/TPAMI.2024.3445770.

A Theoretical Analysis

This section establishes the properties of the teacher composition and calibration rule in Section 3. We fix a student-generated prefix and suppress its index t. The vocabulary $\mathcal{V}$ is finite, and q_L, q_V^I, and q_V^0 are strictly positive distributions, as holds for softmax distributions with finite logits. All logarithms are natural, and $\Delta_{\mathcal{V}}$ denotes the probability simplex over $\mathcal{V}$. We use the convention $0 \log 0 = 0$ in KL divergences.

Define the raw visual residual, source KL, and target budget by

$$r(v) := \log \frac{q_V^I(v)}{q_V^0(v)}, \qquad d := D_{\mathrm{KL}}(q_V^I \| q_V^0), \qquad b := \rho d, \quad \rho \geq 0. \tag{15}$$

For a finite coefficient $\alpha \geq 0$, let

$$\begin{aligned} Z(\alpha) &:= \sum_{v \in \mathcal{V}} q_L(v) e^{\alpha r(v)}, \qquad A(\alpha) := \log Z(\alpha), \\ q_\alpha(v) &:= q_L(v) e^{\alpha r(v) - A(\alpha)}, \quad K(\alpha) := D_{\mathrm{KL}}(q_\alpha \| q_L). \end{aligned} \tag{16}$$

Thus exact calibration requires $K(\alpha^\star) = b$. The results below concern the raw residual and full-vocabulary distributions. Residual stabilization, restricted coefficient statistics, and the sparse student update are described separately in Appendix B.

A.1 Teacher Composition

Proposition 1 (Variational characterization)**.** *For every finite $\alpha \geq 0$, q_α is the unique maximizer of*

$$\max_{q \in \Delta_{\mathcal{V}}} \left\{ \alpha \mathbb{E}_q[r] - D_{\mathrm{KL}}(q \| q_L) \right\}. \tag{17}$$

Proof. For any $q \in \Delta_{\mathcal{V}}$, substituting the definition of q_α gives

$$D_{\mathrm{KL}}(q \| q_\alpha) = D_{\mathrm{KL}}(q \| q_L) - \alpha \mathbb{E}_q[r] + A(\alpha). \tag{18}$$

The objective in Eq. (17) therefore equals $A(\alpha) - D_{\mathrm{KL}}(q \| q_\alpha)$. Nonnegativity of KL, with equality if and only if $q = q_\alpha$, proves the claim. □

Equivalent logit-space form. Let z_L, z_V^I, and z_V^0 be logits for the three source distributions. Each log probability equals its logit minus a branch-specific log-normalizer. Consequently, $\log q_L(v) + \alpha r(v)$ differs from $z_L(v) + \alpha[z_V^I(v) - z_V^0(v)]$ only by a token-independent constant, and

$$q_\alpha = \mathrm{softmax}_{\mathcal{V}}\big(z_L + \alpha[z_V^I - z_V^0]\big). \tag{19}$$

This is an algebraic identity for the source predictions; it does not require the image-free prediction to equal a marginal over images.

A.2 Exact KL Calibration: Existence and Uniqueness

Proposition 2 (Calibration range)**.** *If r is nonconstant, K is continuous and strictly increasing on $[0, \infty)$, with*

$$K(0) = 0, \qquad K'(\alpha) = \alpha \mathrm{Var}_{q_\alpha}[r] > 0 \quad (\alpha > 0). \tag{20}$$

Let $M := \arg\max_{v \in \mathcal{V}} r(v)$ and $q_L(M) := \sum_{v \in M} q_L(v)$. Then

$$\lim_{\alpha \to \infty} K(\alpha) = K_\infty := -\log q_L(M). \tag{21}$$

In particular, $K(\alpha^\star) = b$ has a unique finite positive solution if and only if $0 < b < K_\infty$.

Proof. Differentiating the finite sum defining Z yields

$$A'(\alpha) = \mathbb{E}_{q_\alpha}[r], \qquad A''(\alpha) = \mathrm{Var}_{q_\alpha}[r]. \tag{22}$$

Since $\log(q_\alpha/q_L) = \alpha r - A(\alpha)$,

$$K(\alpha) = \alpha A'(\alpha) - A(\alpha), \qquad K'(\alpha) = \alpha A''(\alpha). \tag{23}$$

For finite α, q_α is strictly positive on $\mathcal{V}$. A nonconstant r therefore has positive variance under q_α, giving strict monotonicity for positive coefficients. Also, $A(0) = 0$ and $q_{\alpha=0} = q_L$, so $K(0) = 0$.

To determine the limiting value, write $r_{\max} := \max_v r(v)$ and subtract it from every residual in the normalized expression for q_α. The factors $e^{\alpha(r(v)-r_{\max})}$ converge to one on M and to zero elsewhere. Hence

$$\lim_{\alpha\to\infty} q_\alpha(v) = \begin{cases} q_L(v)/q_L(M), & v \in M, \\ 0, & v \notin M. \end{cases} \tag{24}$$

On a finite vocabulary with strictly positive q_L, $D_{\mathrm{KL}}(q\|q_L)$ is continuous in q, including at zero coordinates under the stated convention. Applying this continuity to Eq. (24) gives $K_\infty = -\log q_L(M)$. Continuity and strict monotonicity then prove the existence and uniqueness statement. □

For nonconstant r, the zero-budget solution is uniquely $\alpha^\star = 0$. The upper endpoint K_∞ is approached only as $\alpha \to \infty$; no finite coefficient attains it, and larger budgets cannot be attained along this composition family. If r is constant, $q_\alpha = q_L$ and $K(\alpha) = 0$ for every coefficient. For the raw residual in Eq. (15), constancy implies $q_V^I = e^r q_V^0$, and normalization forces $r = 0$ and $d = 0$. We select $\alpha^\star = 0$ in this degenerate case.

KL-constrained interpretation. For $0 < b < K_\infty$, the exactly calibrated teacher also uniquely solves

$$\max_{q\in\Delta_\mathcal{V}} \mathbb{E}_q[r] \quad \text{subject to} \quad D_{\mathrm{KL}}(q\|q_L) \le b. \tag{25}$$

Indeed, Eq. (18) and $K(\alpha^\star) = b$ imply, for every feasible q,

$$\begin{aligned} \alpha^\star\big(\mathbb{E}_q[r] - \mathbb{E}_{q_{\alpha^\star}}[r]\big) &= D_{\mathrm{KL}}(q\|q_L) - b - D_{\mathrm{KL}}(q\|q_{\alpha^\star}) \\ &\le -D_{\mathrm{KL}}(q\|q_{\alpha^\star}) \le 0. \end{aligned} \tag{26}$$

Since $\alpha^\star > 0$, no feasible distribution has a greater expected residual, and equality requires $q = q_{\alpha^\star}$. For $b = 0$, the only feasible distribution is q_L. This characterization holds for a prescribed budget; the source-referenced choice $b = \rho d$ specifies that budget rather than following from constrained optimality.

Finite coefficient caps. For nonconstant r and a finite cap $\alpha_{\max} > 0$, the attainable budgets with $0 \le \alpha \le \alpha_{\max}$ form the closed interval $[0, K(\alpha_{\max})]$. A bounded exact reference can therefore be defined by the unique solution of

$$K(\alpha_{\mathrm{cap}}) = \min\{b, K(\alpha_{\max})\}, \qquad 0 \le \alpha_{\mathrm{cap}} \le \alpha_{\max}. \tag{27}$$

When $b > K(\alpha_{\max})$, this reference returns the cap and falls short of the requested budget. Clipping a locally approximated coefficient need not give this bounded exact solution.

A.3 Local Approximation and Error

For nonconstant r, define $\mu := \mathbb{E}_{q_L}[r]$, $\sigma_L^2 := \mathrm{Var}_{q_L}[r] > 0$, and $\kappa_{3,L} := \mathbb{E}_{q_L}[(r-\mu)^3]$. The finite-vocabulary log-partition function is analytic near zero, with expansion

$$A(\alpha) = \mu\alpha + \tfrac{1}{2}\sigma_L^2\alpha^2 + \tfrac{1}{6}\kappa_{3,L}\alpha^3 + O(\alpha^4), \tag{28}$$

$$K(\alpha) = \tfrac{1}{2}\sigma_L^2\alpha^2 + \tfrac{1}{3}\kappa_{3,L}\alpha^3 + O(\alpha^4). \tag{29}$$

The second identity follows by substituting the first into Eq. (23). Matching its quadratic term to $b = \rho d$ gives the coefficient used in the idealized local rule,

$$\widehat{\alpha} = \sqrt{\frac{2b}{\sigma_L^2}} = \sqrt{\frac{2\rho D_{\mathrm{KL}}(q_V^I\|q_V^0)}{\mathrm{Var}_{q_L}[r]}}. \tag{30}$$

We set $\widehat{\alpha} = 0$ when $\sigma_L^2 = 0$; for the raw residual under the stated assumptions, this also implies $d = 0$.

Local accuracy. To state the approximation precisely, hold q_L and a nonconstant r fixed, introduce an auxiliary budget $0 < \delta < K_\infty$, and let $\alpha^\star(\delta)$ solve $K(\alpha^\star(\delta)) = \delta$. Writing $\widehat{\alpha}(\delta) := \sqrt{2\delta/\sigma_L^2}$, we obtain

$$\alpha^\star(\delta) = \sqrt{\frac{2\delta}{\sigma_L^2}} + O(\delta), \tag{31}$$

$$K(\widehat{\alpha}(\delta)) - \delta = \frac{\kappa_{3,L}}{3}\left(\frac{2\delta}{\sigma_L^2}\right)^{3/2} + O(\delta^2), \qquad \delta \downarrow 0. \tag{32}$$

For the first statement, strict monotonicity implies $\alpha^\star(\delta) \to 0$ as $\delta \downarrow 0$. Eq. (29) then gives $\delta = \frac{1}{2}\sigma_L^2[\alpha^\star(\delta)]^2(1 + O(\alpha^\star(\delta)))$. Taking the nonnegative square root first establishes $\alpha^\star(\delta) = O(\sqrt{\delta})$ and then yields Eq. (31). The second statement follows by substituting $\widehat{\alpha}(\delta)$ into the same expansion.

These are local statements for a fixed prior and residual; their constants may depend on both. Evaluating the approximation at the observed budget $b = \rho d$ does not ensure exact matching or uniformly small error across prefixes. The cubic term also shows that the approximation may overshoot or undershoot the budget. If $b \geq K_\infty$, the formula still defines a finite coefficient when $\sigma_L^2 > 0$, but there is no finite exactly calibrated coefficient to approximate.

Teacher discrepancy along a common composition path. For two finite coefficients $0 \leq a_1 \leq a_2$ with the same prior and residual,

$$\begin{aligned} D_{\mathrm{KL}}(q_{a_2} \| q_{a_1}) &= (a_2 - a_1)A'(a_2) - A(a_2) + A(a_1) \\ &= \int_{a_1}^{a_2} (u - a_1)A''(u)\, du \\ &\leq \int_{a_1}^{a_2} uA''(u)\, du = K(a_2) - K(a_1). \end{aligned} \tag{33}$$

The inequality uses $a_1 \geq 0$ and $A''(u) \geq 0$. If one coefficient is exactly calibrated to b, the right-hand side equals the absolute budget error of the other coefficient. The bound is directional: the KL is taken from the teacher with the larger coefficient to the one with the smaller coefficient. It requires the same prior, residual, and normalized support for both teachers. It therefore does not directly compare a raw-residual teacher with a teacher constructed using a different stabilized residual.

A.4 Special Cases and Invariances

Zero visual budget. If $\rho = 0$, selecting $\alpha = 0$ gives $q_\alpha = q_L$ regardless of the source KL. If $q_V^I = q_V^0$, then $r = 0$ and $d = 0$, and every coefficient yields q_L; our zero-coefficient convention selects a single representative.

Self-prior recovery. Suppose $q_L = q_V^0$ and $\rho = 1$. For the raw residual,

$$q_L(v)e^{r(v)} = q_V^I(v), \qquad Z(1) = 1, \qquad q_{\alpha=1} = q_V^I. \tag{34}$$

Thus $K(1) = D_{\mathrm{KL}}(q_V^I \| q_V^0) = d = b$. If $d > 0$, Proposition 2 implies that the unique exact coefficient is $\alpha^\star = 1$. If $d = 0$, the two source distributions coincide and the zero-coefficient convention recovers the same teacher. This identity concerns exact calibration: the local coefficient $\sqrt{2d/\mathrm{Var}_{q_V^0}[r]}$ need not equal one. Residual modification can also prevent recovery, and a cap below one excludes the exact solution when $d > 0$.

Affine transformations of the residual. For a fixed budget b, let $r'(v) = c\, r(v) + \beta$, where $c > 0$ and β is independent of v. If q'_a denotes composition using r', normalization gives

$$q'_a(v) = \frac{q_L(v)e^{a(cr(v)+\beta)}}{\sum_u q_L(u)e^{a(cr(u)+\beta)}} = q_{ca}(v). \tag{35}$$

A token-independent shift therefore leaves the teacher and its KL displacement unchanged at every coefficient. More generally, whenever the exact matching coefficient is unique, replacing r by r'

changes it to $\alpha^\star/c$ and preserves the calibrated teacher. The unregularized, uncapped local rule has the same property because $\mathrm{Var}_{q_L}[r'] = c^2\mathrm{Var}_{q_L}[r]$. This statement holds with the budget fixed. Changing source-model temperatures can change the source KL itself, while a fixed variance regularizer or coefficient cap can break the coefficient-rescaling property.

A.5 Loss Decomposition

At the fixed prefix, let p_θ and p_θ^0 be the student's strictly positive image-conditioned and image-free distributions, and define $r_\theta^S := \log p_\theta - \log p_\theta^0$. The coefficient α and teacher sources are fixed during the student update. Substitution into the full-vocabulary reverse KL gives

$$\begin{aligned} D_{\mathrm{KL}}(p_\theta \| q_\alpha) &= \mathbb{E}_{p_\theta}[\log p_\theta - \log q_L - \alpha r] + A(\alpha) \\ &= \underbrace{\mathbb{E}_{p_\theta}\left[\log \frac{p_\theta^0}{q_L}\right]}_{\text{language-prior contribution}} + \underbrace{\mathbb{E}_{p_\theta}[r_\theta^S - \alpha r]}_{\text{visual-residual contribution}} + A(\alpha). \end{aligned} \tag{36}$$

This identity holds for any fixed coefficient, including the locally approximated coefficient. The first expectation is taken under p_θ, not p_θ^0, so it is generally not $D_{\mathrm{KL}}(p_\theta^0 \| q_L)$. Neither contribution is individually guaranteed to be nonnegative. The decomposition describes two contributions to a single objective rather than two separately minimized divergences. The image-free student distribution is introduced only for this identity; evaluating the original reverse KL does not require an additional image-free student forward pass.

Because the prefix, teacher sources, and coefficient are fixed, $A(\alpha)$ has zero gradient with respect to θ during the update. Derivatives through $\log p_\theta^0$ cancel between the two contributions when their sum is differentiated consistently. These claims apply to the fixed-prefix objective in Eq. (3), where sampled prefixes are held fixed, and do not differentiate through the rollout sampling distribution.

Relation to the detached visual target. We now relate the visual-residual term in the decomposition above to the detached visual target used in Decomposed OPD (Yoon et al., 2026). Let $\bar\theta$ denote the student parameters at which the target is constructed. The target combines the student's image-free prior $p_{\bar\theta}^0$ with the teacher's visual residual r, with the prior held fixed during differentiation:

$$q_{\mathrm{vis},\bar\theta}(v) := \frac{p_{\bar\theta}^0(v)e^{r(v)}}{\sum_u p_{\bar\theta}^0(u)e^{r(u)}}, \qquad B_{\bar\theta} := \log \sum_u p_{\bar\theta}^0(u)e^{r(u)}. \tag{37}$$

The corresponding KL loss is

$$D_{\mathrm{KL}}(p_\theta \| q_{\mathrm{vis},\bar\theta}) = \mathbb{E}_{p_\theta}[\log p_\theta - \log p_{\bar\theta}^0 - r] + B_{\bar\theta}. \tag{38}$$

At $\theta = \bar\theta$, its value is $\mathbb{E}_{p_{\bar\theta}}[r_{\bar\theta}^S - r] + B_{\bar\theta}$, which is the visual-residual contribution in Eq. (36) at $\alpha = 1$, up to normalization. However, this value correspondence does not by itself imply gradient equivalence. In particular, if $F(\theta) := \mathbb{E}_{p_\theta}[r_\theta^S - r]$ uses an undetached p_θ^0, differentiation yields

$$\begin{aligned} \nabla_\theta D_{\mathrm{KL}}(p_\theta \| q_{\mathrm{vis},\bar\theta})\big|_{\theta=\bar\theta} = {} & \nabla_\theta F(\theta)|_{\theta=\bar\theta} \\ & + \mathbb{E}_{p_{\bar\theta}}\left[\nabla_\theta \log p_\theta^0\big|_{\theta=\bar\theta}\right]. \end{aligned} \tag{39}$$

The extra expectation need not vanish because it is taken under $p_{\bar\theta}$. The target holds the student's prior fixed, whereas $F(\theta)$ differentiates through it. Our reproduction settings and results are reported in Appendix E.3.

B Practical Teacher Construction and Student Update

We describe the numerical safeguards used in visual-residual construction, coefficient estimation, and teacher normalization. All expert branches are frozen and score the same student-generated response prefix over a shared output vocabulary. The image-free branch omits the image and its placeholder while retaining the system prompt, question, and generated text. Below, we suppress the prefix index t and write $q_L = \mathrm{softmax}(z_L)$ and $q_V^c = \mathrm{softmax}(z_V^c)$ for $c \in \{I, 0\}$.

B.1 Visual Residual Stabilization

Ratios involving very small image-free probabilities can amplify errors in tail predictions. We therefore floor low image-free logits, while protecting selected tokens favored by the image. Let $\mathcal{N}_m(q)$ denote the smallest descending-probability nucleus with cumulative mass at least m. The floor is the boundary logit of the original image-free distribution's 0.999-mass nucleus,

$$f := \min_{v \in \mathcal{N}_{0.999}(q_V^0)} z_V^0(v). \tag{40}$$

A token is protected if it lies in the original image-conditioned 0.99-mass nucleus, its probability increases with the image, and its original image-free logit lies below the floor:

$$P := \left\{ v \in \mathcal{N}_{0.99}(q_V^I) : q_V^I(v) > q_V^0(v),\ z_V^0(v) < f \right\}. \tag{41}$$

The effective image-free logits and stabilized residual are

$$\begin{aligned} \widetilde{z}_V^0(v) &:= \begin{cases} z_V^0(v), & v \in P, \\ \max\{z_V^0(v), f\}, & v \notin P, \end{cases} \\ \widetilde{r}(v) &:= z_V^I(v) - \widetilde{z}_V^0(v). \end{aligned} \tag{42}$$

Flooring reduces the residual for unprotected tokens with very low image-free probability. Protection preserves the original contrast for tokens satisfying all three conditions in Eq. (41).

We then center the residual by subtracting its maximum,

$$u(v) := \widetilde{r}(v) - \max_{a \in \mathcal{V}} \widetilde{r}(a). \tag{43}$$

This constant shift preserves the residual variance and the normalized teacher, while keeping the transferred residual nonpositive. The original branch probabilities are retained for support selection and source-KL estimation; flooring is applied only when constructing the transferred residual.

B.2 Calibration Coefficient Estimation

To estimate the coefficient from high-probability candidates, we form the union of the three original teacher nuclei,

$$S := \mathcal{N}_m(q_L) \cup \mathcal{N}_m(q_V^I) \cup \mathcal{N}_m(q_V^0). \tag{44}$$

We use a calibration support mass of $m = 0.95$; Appendix D.4 compares it with $m = 0.99$. Each branch is renormalized on the common set S: for $q \in \{q_L, q_V^I, q_V^0\}$, define $q^S(v) := q(v) / \sum_{a \in S} q(a)$ for $v \in S$. The coefficient uses the original image/no-image contrast and the variance of the stabilized residual under the receiving language prior,

$$\begin{aligned} d_S &:= D_{\mathrm{KL}}(q_V^{I,S} \| q_V^{0,S}), \\ V_S &:= \mathrm{Var}_{q_L^S}[\widetilde{r}]. \end{aligned} \tag{45}$$

The union retains candidates assigned high probability by any of the three branches. For every branch q, $q(S) \geq m$ and $\mathrm{TV}(q, q^S) = 1 - q(S) \leq 1 - m$, where q^S is extended by zero outside S. Thus m controls the excluded model probability mass. We use d_S as a conditional calibration statistic to reduce dependence on low-probability likelihood ratios; high retained mass alone does not guarantee closeness to the full-vocabulary source KL.

We use $\rho = 1$ by default, regularize the denominator, and cap the coefficient:

$$\widehat{\alpha}_S := \min\left\{ \alpha_{\max}, \sqrt{\frac{2\rho d_S}{V_S + \epsilon}} \right\}, \qquad \epsilon = 10^{-8}, \quad \alpha_{\max} = 6. \tag{46}$$

The regularizer makes the expression defined when the estimated variance is zero, and the cap limits amplification when that variance is small. This cap is a practical safeguard against large coefficients, not a way to make an unattainable KL target attainable. We use the same $\alpha_{\max} = 6$ across adaptive training variants. A zero estimated source KL gives a zero coefficient. We compute the coefficient statistics in FP32 and clamp numerical KL and variance estimates to be nonnegative.

As a numerical safeguard, NaN and negative-infinite coefficients are replaced by zero and positive-infinite coefficients by $\alpha_{\max}$ before clamping to $[0, \alpha_{\max}]$.

This is a regularized, capped second-order estimate, rather than an exact solution of the KL-budget equation. Support restriction and residual stabilization also affect the approximation, so the practical teacher need not attain the full-vocabulary displacement $\rho D_{\mathrm{KL}}(q_V^I \| q_V^0)$. Its induced KL displacement is evaluated in Appendix D.5.

B.3 Teacher Normalization and Student Update

The final teacher is normalized over the entire output vocabulary,

$$q_T(v) := \frac{\exp\{z_L(v) + \widehat{\alpha}_S u(v)\}}{\sum_{a \in \mathcal{V}} \exp\{z_L(a) + \widehat{\alpha}_S u(a)\}}. \tag{47}$$

Thus S restricts coefficient estimation, not teacher support.

The variational characterization in Appendix A.1 applies to any finite evidence score. Consequently, the practical teacher in Eq. (47) uniquely maximizes $\widehat{\alpha}_S \mathbb{E}_q[u] - D_{\mathrm{KL}}(q \| q_L)$ over $q \in \Delta_{\mathcal{V}}$. Stabilization changes the transferred score and its coefficient while preserving this generalized Bayesian form.

We evaluate the full-vocabulary log-normalizer using FP32 log-sum-exp reductions over vocabulary chunks, combined with log-add-exp. This preserves the full normalization without materializing a full FP32 tensor of composed teacher logits. The distillation temperature is 1.

Adopted top-16 OPD update. We follow the student-top-16 OPD setting of Li et al. (2026), using the composed q_T as the teacher. For reproducibility, we specify the support and gradient placement. Let $\bar{\theta}$ denote the current scoring parameters and let p_θ be the image-conditioned student distribution at the fixed prefix. We select $k = 16$ tokens using detached student scores,

$$U := \mathrm{TopK}_{v \in \mathcal{V}}\, p_{\bar{\theta}}(v), \qquad w(v) := \frac{p_{\bar{\theta}}(v)}{\sum_{a \in U} p_{\bar{\theta}}(a)}, \quad v \in U. \tag{48}$$

Only the weights are renormalized on U. The student and teacher log probabilities retain their full-vocabulary normalizers. The token-level surrogate defining the update is

$$\widehat{\mathcal{L}}(\theta) := \sum_{v \in U} \mathrm{sg}\left[w(v) \log \frac{p_{\bar{\theta}}(v)}{q_T(v)} \right] \log p_\theta(v). \tag{49}$$

The weights and log ratios are recomputed at the current scoring parameters and detached before differentiation. Gradients flow only through the final student log-probability factor; the prefix, expert predictions, teacher target, and selected support are fixed.

The teacher-derived support S is used only for coefficient estimation, whereas the student-derived support U selects the terms in this update. Both student and teacher log probabilities remain normalized over $\mathcal{V}$.

C Experimental Details

C.1 Models and Training

Table 5 summarizes the reference training configuration and the sampled-decoding settings for reasoning evaluation. Perception benchmarks use the separate settings in Appendix C.3.2. Baseline-specific adaptations are described in Appendix C.5.

Table 5: Training and evaluation settings used in our experiments. The training configuration corresponds to the main LEGO-OPD setup and serves as the reference configuration for the baseline comparisons. Baselines retain these settings whenever compatible with their defining formulations, with method-specific deviations described in Appendix C.5. The reasoning-evaluation settings apply to sampled decoding and are held fixed across compared methods. Perception settings are given in Appendix C.3.2.

Category	Hyperparameter	Value
Models	Student	Qwen3-VL-2B-Instruct
	Language expert	Qwen3-4B-Instruct-2507
	Grounding expert	Qwen3-VL-4B-Instruct
Training	Training data	MMFineReason, 54,840 examples
	Student update	Full-parameter fine-tuning
	Training steps	200
	Responses per update	128
	Optimizer	AdamW
	Learning rate	1×10^{-6}
	Weight decay	0.01
	LR schedule	Constant, no warmup
Training rollout	Maximum prompt length	12,384 tokens
	Maximum response length	12,384 tokens
	Temperature	1.0
	Top-p	1.0
	Top-k	Unrestricted
	Min-p	0
	Repetition penalty	1.0
LEGO-OPD	OPD support	Student top 16
	KL-budget scale	$\rho = 1$
	Calibration support mass	0.95
	Coefficient cap	$\alpha_{\max} = 6$
	Null-floor mass	0.999
	Image-protection mass	0.99
Reasoning eval.	Temperature	0.9
	Top-p	0.95
	Top-k	100
	Min-p	0.01
	Repetition penalty	1.03
	Maximum response length	12,384 tokens
	Stop sequence	`</answer>`
	Inference repetitions	5 (seeds 42–46)

Training. We fully fine-tune the student while keeping the language and grounding experts frozen. At each update, the current student generates on-policy responses, which are then scored by the expert branches. We do not use teacher-generated responses, ground-truth answers, or reference solutions as training targets.

We use the prompt template shown in Figure 3 throughout training. The language expert receives the textual problem only, while the grounding expert is evaluated with and without the image under

the same generated response prefix. LEGO-OPD uses the student-top-16 surrogate together with the practical calibration and residual stabilization in Appendix B.

C.2 Benchmarks and Metrics

We evaluate multimodal reasoning on MATH-Vision, MathVerse, MathVista, MMMU-Pro, RBench-M, and the multimodal split of OlympiadBench, and text-only reasoning on MATH-500, MathVerse Text Only, MMLU-Pro, RBench-T, and the text-only split of OlympiadBench. Visual perception is evaluated on BLINK (Fu et al., 2025) and QBench1,2 (Wu et al., 2024; Zhang et al., 2024). For QBench1,2, we use the perception-focused Group A questions, namely the A1 perception questions from the development splits (1,495 from QBench 1 and 1,000 from QBench 2), and report accuracy over the pooled 2,495 questions. The six multimodal and five text-only reasoning scores are averaged within their respective groups. BLINK and QBench1,2 are reported separately and excluded from both reasoning averages.

C.3 Evaluation Protocol

Reasoning and perception benchmarks follow separate protocols. Reasoning benchmarks use sampled decoding with the evaluation settings in Table 5 (Appendix C.3.1). Perception benchmarks use greedy direct-answer generation with the settings in Table 6 (Appendix C.3.2).

C.3.1 Reasoning Benchmarks

Inputs. All reasoning benchmarks except MMLU-Pro use the shared system prompt in Figure 3. The user prompt contains each question in its benchmark's native format, together with any benchmark-specific instructions. MathVerse and R-Bench add their own instructions to the question, and OlympiadBench uses its multimodal or text-only prompt template. MMLU-Pro uses category-conditioned five-shot prompts without a system prompt.

Decoding and reporting. We sample five responses per question with evaluation seeds 42–46 under the reasoning-evaluation settings in Table 5. The main text reports the mean accuracy over these samples, and Appendix E reports the same results with standard deviations (mean $\pm$ SD). Appendix E.1.2 additionally reports greedy-decoding results. These SDs reflect sampling variability across evaluation seeds for each trained checkpoint. Benchmark averages are unweighted means and are reported as means only.

Answer extraction and scoring. For multiple-choice tasks, predictions are first mapped to answer options by deterministic rules. Only when this mapping remains unresolved do we use `gpt-5.6-luna` as a gold-blind semantic option mapper, with the prompt in Figure 4. The mapper receives only the extracted final answer and the candidate options.

MMLU-Pro is scored by a dedicated deterministic A–J evaluator. OlympiadBench uses its benchmark-specific mathematical answer normalization and automatic scorer, and MathVista and MATH-500 use their task-specific answer normalization. MathVerse uses its dedicated answer extraction and correctness-judging pipeline, calling `gpt-5.6-luna` when an LLM-based judgment is required.

C.3.2 Perception Benchmarks

Inputs and decoding. BLINK uses its dataset-provided prompt, and QBench1,2 uses the question followed by its A–D choices. We append the instruction listed in Table 6, use the Qwen3-VL processor and chat template without a system prompt, and generate one response per question with greedy decoding (temperature 0) and seed 42.

Scoring. The option letter is extracted from each response by deterministic rules and compared with the gold letter by exact match. Responses without an extractable letter are counted as incorrect. For BLINK, the letter is taken from explicit answer patterns or a standalone final choice, whereas for QBench1,2 it is taken from the beginning of the response. BLINK accuracy is computed over all 1,901 validation questions, and QBench1,2 accuracy over the 2,495 pooled Group A questions.

Table 6: Settings for direct-answer perception evaluation. Both benchmarks use one greedy response per question (temperature 0, seed 42) and no system prompt. Max. tokens is the maximum number of generated tokens.

Benchmark	Questions	Max. tokens	Appended instruction	Scoring
BLINK validation	1,901	128	"Reply with only the letter of the correct choice (A, B, C, ...). Do not explain."	Rule-based letter extraction and exact match, accuracy over all questions
QBench1,2 Group A (A1 perception, dev)	2,495 (1,495 + 1,000)	16	"Reply with only the letter of the correct choice (A, B, C, or D). Do not explain."	Leading-letter extraction and exact match, accuracy over all pooled questions

C.4 Prompt Templates

Figures 3 and 4 give the QA template used for training and reasoning evaluation and the semantic option-mapping prompt. Benchmark-specific formats and the separate perception settings are described in Appendices C.3.1 and C.3.2.

Common QA Prompt

System Prompt: Please reason step by step, and put your final answer within `<answer>\boxed{your final answer}</answer>`.

User Prompt: {Question and answer choices}

Figure 3: Common QA prompt used for training and reasoning evaluation. Benchmark-specific input formatting is preserved during evaluation. MMLU-Pro uses its native five-shot prompt without the shared system prompt. Perception settings are given in Appendix C.3.2.

Semantic Option-Mapping Prompt for Scoring

User Prompt: Map the provided final answer to an option, or null. This is ONLY answer interpretation, not solving the original question. No correct answer key is supplied.
Rules:
- A clear standalone option letter selects that option.
- A clear final value or description can select the unique semantically equivalent option.
- If BOTH a letter and answer content are stated, they MUST agree with that option's content. Otherwise output null. Example: B:52 is null when B contains 64, even if another option contains 52.
- A letter plus reasoning can be interpreted only when the stated answer content is consistent; do not evaluate unrelated intermediate reasoning or solve a question.
- Ignore Markdown, answer tags, boxing, harmless punctuation, and equivalent mathematical notation.
- If the supplied final regions conflict, contain competing answers, or cannot be mapped with certainty, output null.
- Do not infer an option from mathematical knowledge of what the answer SHOULD be. Return ONLY a JSON object {"option": "A"} (using the selected letter) or {"option": null}.

Data: {options and model output}

Figure 4: Prompt used by GPT-5.6 Luna for gold-blind semantic option mapping during scoring. The judge receives only the extracted final answer and candidate options, without the original question, image, or reference answer.

C.5 Baseline Implementations

We use the LEGO-OPD training setup in Table 5 whenever compatible with each baseline. Unless otherwise noted, we retain the objectives and method-specific hyperparameters reported in the original papers, while adapting the model configuration, data, and training budget to our experimental setting.

Standard OPD. Standard OPD distills directly from a frozen Qwen3-VL-4B-Instruct teacher conditioned on the image and question. We use the same MMFineReason data, rollout configuration, and student-top-16 OPD surrogate as LEGO-OPD.

VA-OPD. VA-OPD retains its multi-response training scheme, generating four responses per prompt and therefore using 32 unique prompts per update under the common 128-response budget. Following the reported configuration, the VLM teacher scores each response with the original and pixelated image to estimate visual importance. We use a degradation ratio of 0.1, assign the highest-advantage 20% of token positions to the high-importance group, and give equal weight to the high- and low-importance groups. Training uses full-vocabulary reverse KL rather than the student-top-16 surrogate.

M-OPD. M-OPD was originally formulated for domain-routed multi-teacher distillation over text domains. We adapt it to the multimodal setting by routing examples from our multimodal training data (the 54,840-example MMFineReason subset in Section 4.1) to the VLM teacher and DeepMath-103K (He et al., 2026) examples to the LLM teacher at a $1{:}1$ ratio. Following M-OPD's policy-gradient formulation, we use sampled-token distillation with the reported advantage clipping magnitude of 5 instead of a top-K objective.

H-OPD. H-OPD combines a VLM teacher conditioned on the image and question with an LLM teacher conditioned on the released description-augmented input. Following its reported configuration, each teacher distribution is normalized over its top-eight tokens, and the two distributions are combined using entropy-based confidence weights with temperature 0.5. The student is trained with reverse KL over the union of the two teacher supports. This support contains at most 16 tokens and is distinct from the student-top-16 support used by Standard OPD and LEGO-OPD.

D Analysis of Source-Referenced Calibration

This section analyzes the calibration rule in three parts. Appendix D.1 uses a three-token example to show why a fixed coefficient fails when the language prior differs from the grounding expert's image-free prior. Appendices D.2, D.3, and D.4 vary one component of the calibration in training: the coefficient rule and KL-budget scale, the zero-budget limit, and the calibration support mass. Appendices D.5 and D.6 evaluate the practical calibration rule, first against numerical KL matching and then through distillation from a recomposed VLM teacher.

D.1 A Three-Token Calibration Example

Section 3 calibrates the evidence scale so that the composed teacher's displacement from the language prior matches the scaled source KL in Eq. (11). The visual residual is measured against the grounding expert's own image-free prediction q_V^0 but applied to the language expert's prior q_L. When the two priors coincide, $q_L = q_V^0$, exact calibration gives $\alpha^\star = 1$ and recovers q_V^I Eq. (34) (Appendix A.4). We use a three-token example in which the LLM prior differs substantially from the VLM's image-free prior to show why a single fixed coefficient cannot meet the calibration reference across prefixes (Figure 5). Each row fixes a prefix. The grounding expert provides q_V^0 and q_V^I, the language expert provides q_L, and the composed teacher is $q_\alpha(v) \propto q_L(v)\exp\{\alpha r(v)\}$ as in Eq. (9). We use $\rho = 1$, so exact calibration requires $D_{\mathrm{KL}}(q_{\alpha^\star}\|q_L) = D_{\mathrm{KL}}(q_V^I\|q_V^0)$.

Reasoning prefix. The top row of Figure 5 shows a prefix at which the image is nearly irrelevant. The grounding expert changes only at a low-mass token, where $q_V(v_1)$ moves from 0.01 to 0.02, and its argmax stays at v_2. The source KL is therefore small, $D_{\mathrm{KL}}(q_V^I\|q_V^0) = 0.004$. The residual $r(v_1) = \log 2$ is nevertheless large, because it is a ratio. With $\alpha = 1$, this ratio is applied to a language prior that already assigns 0.35 to v_1. The teacher's probability of v_1 rises to 0.52, which overturns the language expert's preferred token v_2. The resulting displacement is $D_{\mathrm{KL}}(q_{\alpha=1}\|q_L) =$

0.062, or 15.7 times the source KL. Calibration selects $\alpha^\star = 0.26$, which matches the source KL and preserves v_2.

Grounding prefix. The bottom row of Figure 5 shows a prefix at which the image determines the answer. The grounding expert moves from v_2 to v_1, with $q_V(v_1)$ increasing from 0.20 to 0.80, which gives $D_{\mathrm{KL}}(q_V^I \| q_V^0) = 0.832$. The language expert cannot see the image and assigns only 0.03 to v_1. With $\alpha = 1$, the teacher moves by only $D_{\mathrm{KL}}(q_{\alpha=1} \| q_L) = 0.546$, or 0.66 times the source KL, and still prefers v_2. Calibration selects $\alpha^\star = 1.15$, and the teacher follows the image.

Discussion. The two prefixes require coefficients that differ by more than a factor of four, so no single fixed value suits both. Raising the fixed coefficient to $\alpha = 2$ corrects the grounding prefix ($q_\alpha(v_1) = 0.89$), but it increases the displacement at the reasoning prefix to 60 times the source KL ($q_\alpha(v_1) = 0.69$). This pattern mirrors Table 2, where larger fixed coefficients improve perception but reduce text-only reasoning. Source-referenced calibration avoids this trade-off by tying the displacement at each prefix to how much the image changed the grounding expert itself.

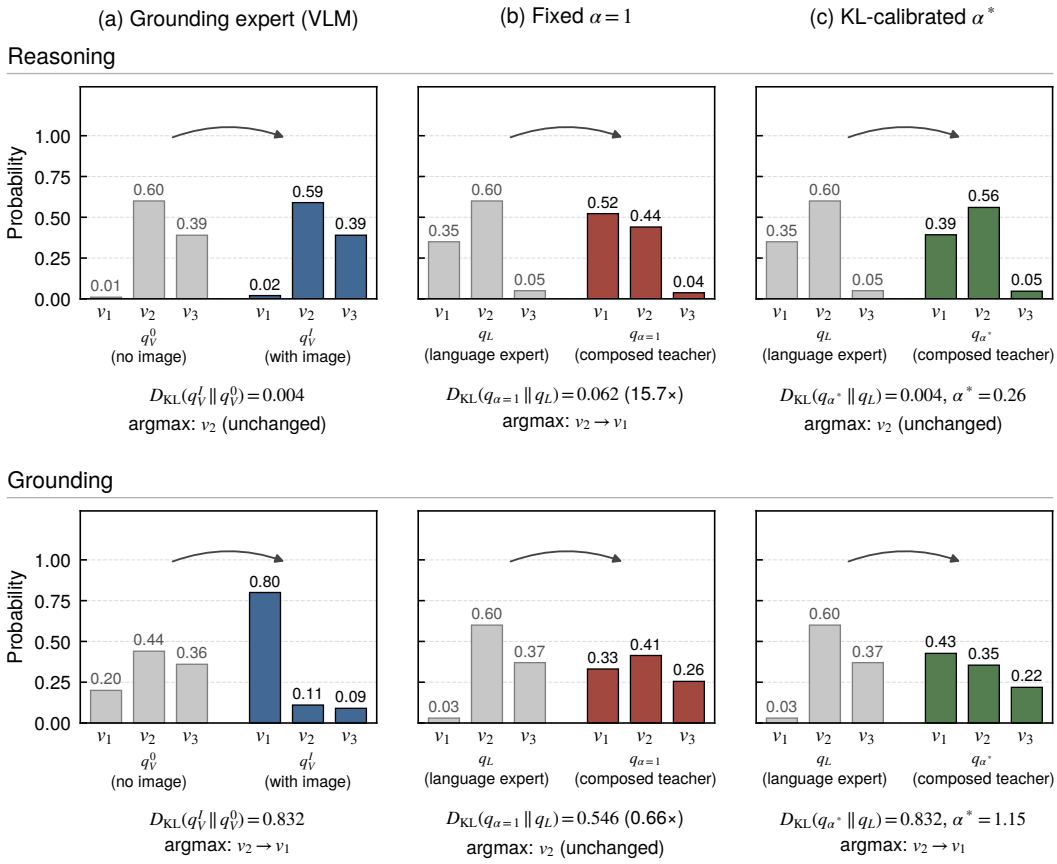


Figure 5: **Illustrative example: a fixed coefficient cannot fit every prefix.** Gray bars show the next-token distribution before the update (q_V^0 or q_L), and colored bars show it after the update (q_V^I or the composed teacher). *Top:* at a reasoning prefix, the image barely changes the grounding expert ($D_{\mathrm{KL}}(q_V^I \| q_V^0) = 0.004$). With $\alpha = 1$, however, the teacher moves 15.7× the source KL, and its argmax changes from v_2 to v_1. *Bottom:* at a grounding prefix, the image determines the answer ($D_{\mathrm{KL}}(q_V^I \| q_V^0) = 0.832$). With $\alpha = 1$, the teacher reaches only 0.66× the source KL and keeps v_2. KL calibration with $\rho = 1$ selects $\alpha^\star = 0.26$ and 1.15, respectively, which match the source KL in both cases. Multiples in parentheses are relative to the source KL.

D.2 Fixed and Adaptive Calibration

Fixed-coefficient ablation. We compare adaptive calibration at $\rho = 1$ with constant coefficients $a \in \{1, 2, 4\}$ using the same student and the same 4B language and grounding experts as in Table 5. The only change is to replace the prefix-dependent coefficient $\widehat{\alpha}_{S,t}$ with a at every response position:

$$q_{T,t}^{\text{fixed}} := \text{softmax}_{\mathcal{V}}\big(z_{L,t} + au_t\big). \tag{50}$$

All runs retain the residual stabilization, teacher normalization, student-top-16 update, and training and evaluation settings in Appendices B and C.1. Only coefficient estimation is bypassed in the fixed variants.

KL-budget scale. For the adaptive variants, we compare $\rho \in \{1, 2, 4\}$ while retaining the remaining settings in Table 5. All variants use the practical coefficient estimator in Eq. (46), with calibration support mass $m = 0.95$, variance regularizer $\epsilon = 10^{-8}$, and coefficient cap $\alpha_{\max} = 6$. Changing ρ rescales the source-KL budget before coefficient estimation. Results for all reasoning and perception benchmarks are reported in Table 2. Table 7 adds evaluation SDs for all reasoning results in Table 2, using the reporting protocol in Appendix C.3.

Table 7: **Calibration ablations with evaluation variability.** All runs use the same 2B student and 4B language and grounding experts for 200 training steps. Fixed composition uses constant α, while adaptive calibration uses KL-budget scale ρ. Scores are mean $\pm$ SD over five evaluation sampling seeds. Averages report means only.

	Fixed coefficient			**Adaptive calibration**		
Benchmark	$\alpha = 1$	$\alpha = 2$	$\alpha = 4$	$\rho = 1$	$\rho = 2$	$\rho = 4$
Multimodal Reasoning						
MATH-Vision	31.91 ± 1.09	31.25 ± 2.58	30.00 ± 1.42	33.36 ± 2.56	34.28 ± 2.06	34.01 ± 0.76
MathVerse (Total)	53.46 ± 0.44	54.29 ± 0.50	53.41 ± 0.69	56.96 ± 0.20	57.19 ± 0.44	57.51 ± 0.51
MathVista	64.58 ± 0.72	68.98 ± 0.79	69.36 ± 0.33	69.62 ± 0.72	70.58 ± 1.02	70.50 ± 0.55
MMMU-Pro	39.48 ± 0.77	38.87 ± 0.66	37.45 ± 0.64	40.98 ± 1.14	41.38 ± 0.72	41.36 ± 0.61
RBench-M	26.89 ± 1.63	24.03 ± 0.98	17.62 ± 0.78	29.17 ± 1.20	29.20 ± 1.49	28.39 ± 2.51
OlympiadBench-MM	38.27 ± 2.03	41.87 ± 2.72	41.73 ± 3.70	44.93 ± 2.24	45.07 ± 3.85	45.33 ± 1.56
Multimodal Avg.	42.43	43.22	41.60	45.84	46.28	46.18
Text Reasoning						
MATH-500	82.24 ± 1.10	82.08 ± 1.12	81.52 ± 1.54	82.40 ± 0.99	82.36 ± 0.84	82.52 ± 0.88
MathVerse (Text Only)	59.06 ± 0.72	58.63 ± 1.14	57.94 ± 0.68	59.59 ± 0.99	60.23 ± 0.40	60.46 ± 0.50
MMLU-Pro	47.40 ± 0.47	47.16 ± 0.25	47.10 ± 0.10	47.15 ± 0.05	47.54 ± 0.18	47.38 ± 0.26
RBench-T	38.21 ± 1.94	35.52 ± 0.70	32.85 ± 0.64	38.59 ± 1.19	40.27 ± 1.99	38.08 ± 0.59
OlympiadBench-Text	52.20 ± 1.33	51.72 ± 1.42	51.19 ± 1.20	52.31 ± 0.89	53.00 ± 0.77	51.78 ± 1.09
Text Avg.	55.82	55.02	54.12	56.01	56.68	56.04

D.3 Language-Only Supervision at Zero KL Budget

The dataset released by H-OPD (Yin et al., 2026) provides image-and-question inputs for the VLM and question-and-image-description inputs for the LLM. We refer to these formats as *multimodal data* and *text-only data*, respectively.

To examine the role of grounding supervision, we compare two language-only controls at $\rho = 0$, where the teacher reduces to the language expert, $q_{T,t} = q_{L,t}$. One control uses multimodal data, matching the data configuration of the $\rho = 1$ runs. The other uses text-only data, giving the LLM teacher access to image descriptions. Apart from ρ and the data format, both controls follow the $\rho = 1$ training and evaluation settings, including the same 2B student initialization, 4B language expert, and 200 training steps. Table 8 compares these controls with the initial student and two $\rho = 1$ configurations from Table 4. Both configurations use multimodal data and the same 4B language expert, with either the frozen initial student or the 4B VLM providing the visual residual.

Table 8: **Language-only supervision and expert composition.** Base 2B is the initial student before distillation. Student (2B) denotes its frozen copy used as the grounding expert. Scores are accuracies (%), reported as mean ± SD over five evaluation sampling seeds. Benchmark averages report means only.

	Base 2B	**Distilled 2B students**			
Training data	–	**Multimodal**	**Text-only**	**Multimodal**	**Multimodal**
Grounding expert	–	–	–	Student (2B)	VLM 4B
KL-budget scale	–	$\rho = 0$	$\rho = 0$	$\rho = 1$	$\rho = 1$
Multimodal Reasoning					
MATH-Vision	27.96 ± 1.68	21.05 ± 2.17	33.75 ± 1.80	33.03 ± 1.35	33.36 ± 2.56
MathVerse (Total)	51.51 ± 0.48	30.25 ± 0.18	54.86 ± 0.53	54.17 ± 0.65	56.96 ± 0.20
MathVista	66.56 ± 0.71	32.22 ± 0.74	66.34 ± 0.75	69.42 ± 1.13	69.62 ± 0.72
MMMU-Pro	37.40 ± 0.66	23.93 ± 0.14	39.58 ± 0.57	41.93 ± 0.69	40.98 ± 1.14
RBench-M	19.61 ± 0.95	25.86 ± 0.66	26.95 ± 0.65	28.48 ± 1.95	29.17 ± 1.20
OlympiadBench-MM	37.60 ± 3.42	32.80 ± 1.28	39.73 ± 1.67	43.07 ± 3.55	44.93 ± 2.24
Multimodal Avg.	40.11	27.69	43.54	45.02	45.84
Text Reasoning					
MATH-500	80.44 ± 0.95	82.44 ± 0.89	82.08 ± 0.97	82.36 ± 0.57	82.40 ± 0.99
MathVerse (Text Only)	57.18 ± 1.01	60.46 ± 0.74	59.57 ± 0.86	59.92 ± 0.61	59.59 ± 0.99
MMLU-Pro	43.59 ± 0.17	48.11 ± 0.33	47.44 ± 0.44	47.51 ± 0.35	47.15 ± 0.05
RBench-T	30.20 ± 0.31	38.21 ± 1.51	40.05 ± 0.59	38.46 ± 1.30	38.59 ± 1.19
OlympiadBench-Text	48.04 ± 1.42	52.61 ± 0.83	52.73 ± 0.58	52.67 ± 1.35	52.31 ± 0.89
Text Avg.	51.89	56.37	56.37	56.18	56.01

These results suggest that text-only reasoning gains do not by themselves ensure effective multimodal transfer. The benefit of incorporating the frozen student's visual residual supports a complementary role for image-conditioned supervision. In this setting, factor composition can therefore benefit from the student's existing visual signal without requiring a stronger grounding model.

D.4 Sensitivity to Calibration Support Mass

To assess sensitivity to the calibration support, we compare nucleus masses $m \in \{0.95, 0.99\}$ in the union construction of Eq. (44). Both runs use the 4B language and grounding experts, the same 2B student initialization, $\rho = 1$, and 200 training steps. All other training, teacher-construction, and evaluation settings are held fixed. Only the mass threshold used to construct the coefficient-estimation support S is changed; the student top-16 support rule and full-vocabulary teacher normalization are unchanged.

Figure 6 shows the training diagnostics. Both settings produce broadly similar decreases in training and top-k OPD losses. The entropy-gap and support-mass diagnostics also follow similar trajectories, with differences in the top-k KL estimate at several stages of training.

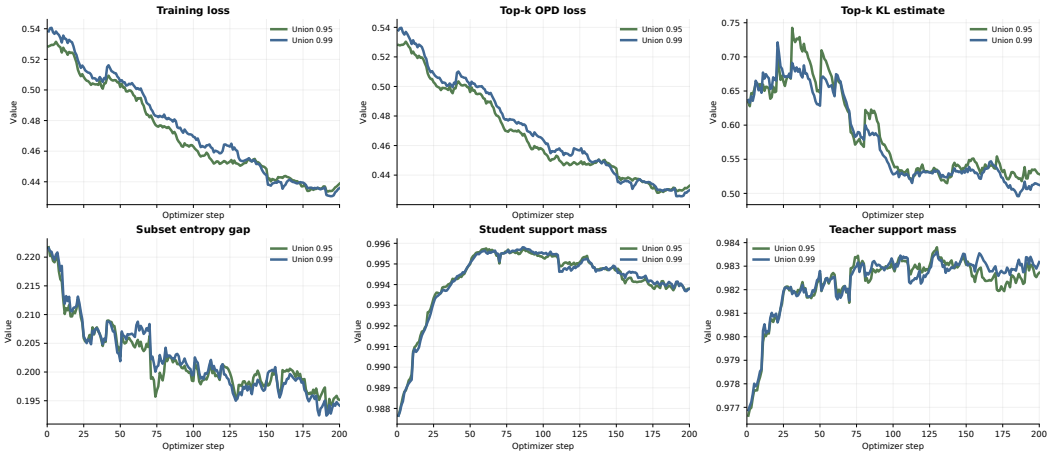


Figure 6: **Training diagnostics for calibration support masses 0.95 and 0.99.** The 4B–4B composition is trained at $\rho = 1$ with all other settings fixed. Panels compare training loss, top-k OPD loss, top-k KL estimate, subset entropy gap, student support mass, and teacher support mass over 200 optimizer steps.

Table 9: **Effect of calibration support mass on final reasoning performance.** Both runs use the same 2B student and 4B language and grounding experts for 200 steps at $\rho = 1$. Accuracies (%) are mean $\pm$ SD across five evaluation sampling seeds. Averages are unweighted means of the benchmark means.

Benchmark	$m = 0.95$ (default)	$m = 0.99$
	Multimodal Reasoning	
MATH-Vision	33.36 ± 2.56	35.20 ± 1.45
MathVerse (Total)	56.96 ± 0.20	56.97 ± 0.54
MathVista	69.62 ± 0.72	69.62 ± 0.51
MMMU-Pro	40.98 ± 1.14	40.74 ± 0.47
RBench-M	29.17 ± 1.20	28.39 ± 0.79
OlympiadBench-MM	44.93 ± 2.24	44.67 ± 2.45
Multimodal Avg.	45.84	45.93
	Text Reasoning	
MATH-500	82.40 ± 0.99	81.80 ± 0.65
MathVerse (Text Only)	59.59 ± 0.99	59.80 ± 0.91
MMLU-Pro	47.15 ± 0.05	47.51 ± 0.29
RBench-T	38.59 ± 1.19	38.90 ± 1.05
OlympiadBench-Text	52.31 ± 0.89	52.11 ± 0.71
Text Avg.	56.01	56.02

Table 9 shows mixed changes across individual benchmarks, with similar final averages: 45.84 versus 45.93 for multimodal reasoning and 56.01 versus 56.02 for text-only reasoning at $m = 0.95$ and $m = 0.99$, respectively. Together with the training curves, these results indicate limited sensitivity to increasing the support mass from 0.95 to 0.99 in this configuration and support retaining $m = 0.95$ as the default.

D.5 Approximation Accuracy of the Practical Coefficient

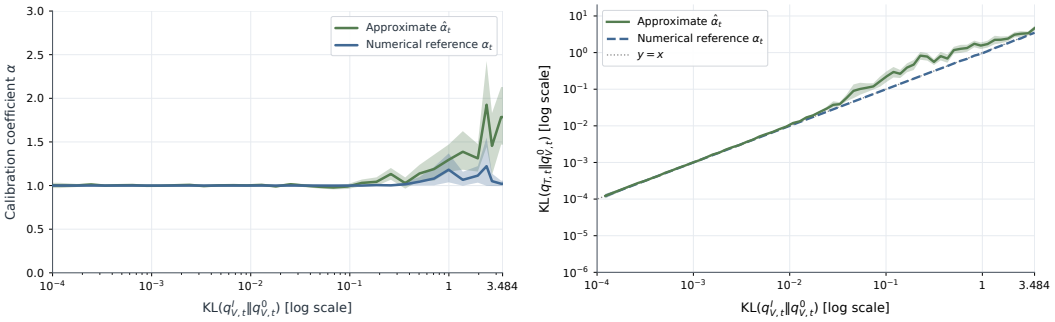
Training approximates the exact KL matching in Eq. (11) with the second-order coefficient in Eq. (13), computed on a restricted support with a variance regularizer and a coefficient cap as in Eq. (46). We measure how closely this practical coefficient matches a numerical solution of the KL-matching equation on prefixes generated during training, in two settings. Figure 2 in the main text composes the 4B language and grounding experts, with the language expert's distribution as the receiving prior $q_{L,t}$. Figure 7 uses self-prior recomposition, which instead sets $q_{L,t} = q^0_{V,t}$ with the grounding VLM's own image-free prediction.

Setup. Both settings use $\rho = 1$ and the same calibration support and residual stabilization as in student training. At each prefix, we construct the stabilized residual u_t as in Appendix B.1 and the support S_t as in Eq. (44) with $m = 0.95$. For expert composition, S_t is the union of the language, image-conditioned, and image-free nuclei. For self-prior recomposition, the language and image-free branches coincide, leaving two distinct nuclei. All distributions are renormalized on S_t before computing KL. The source KL uses the original image-conditioned and image-free predictions, and the teacher KL measures the displacement of the composed teacher from the receiving prior on S_t. Distillation itself uses the full-vocabulary teacher.

Compared coefficients. The practical coefficient $\widehat{\alpha}_{S,t}$ is computed by Eq. (46), including the variance regularizer and coefficient cap. The numerical reference solves for the coefficient at which the teacher KL equals the source KL, using the same residual, receiving prior, and support. In Figures 2 and 7, panel (a) compares the two coefficients, and panel (b) compares the resulting teacher KL with the source KL, where the diagonal marks exact matching.

Relation to exact self-prior recovery. With the raw residual, composing $q^0_{V,t}$ at $\alpha = 1$ recovers $q^I_{V,t}$, also after renormalization on a common support. Residual stabilization changes the composition path, so the numerical reference in Figure 7 can differ from one. It therefore serves as a reference for the practical self-prior path. Appendix D.6 evaluates the effect of this practical construction on distillation.

Prefixes and aggregation. The diagnostic uses 24,576 prefixes from 768 rollouts on 256 MMMU (Yue et al., 2024) questions, generated by student checkpoints at steps 0, 100, and 200. MMMU is separate from the MMMU-Pro benchmark used for evaluation. The practical and numerical coefficients are evaluated on the same prefixes. The figures report means within source-KL bins, with pointwise 95% confidence intervals from question-level bootstrap resampling. Teacher-KL curves average per-prefix divergences. Binned means can agree even when individual prefixes deviate, and the upward deviations at larger source KL reflect finite-update error in the practical approximation.



(a) Calibration coefficient (b) Induced teacher KL

Figure 7: **Local approximation error in self-prior recomposition.** The 4B VLM supplies both its own image-free prior and visual residual, with $\rho = 1$. Panels compare calibration coefficients (a) and teacher KL relative to the VLM's image-free prior (b). Curves show means within source-KL bins. The dotted $y = x$ line denotes source-KL matching. Shaded bands indicate question-bootstrap 95% confidence intervals.

D.6 Distillation with a Recomposed VLM Teacher

We define a *recomposed teacher* by decomposing a single VLM into its own image-free prior and visual residual and recombining them with adaptive calibration at $\rho = 1$. The receiving prior is the VLM's image-free distribution, $q_{L,t} := q^0_{V,t}$, and its image-conditioned and image-free predictions supply the visual residual. The recomposed teacher thus replaces the external language prior of the main composition with the grounding expert's own prior. To validate the practical implementation, we compare a student trained with this teacher against standard OPD, which distills directly from the frozen Qwen3-VL-4B-Instruct teacher.

All other settings match the main experiment, including the student initialization, training data and budget, evaluation protocol, and the practical construction in Appendix B (residual stabilization, $m = 0.95$, $\epsilon = 10^{-8}$, and $\alpha_{\max} = 6$). The recomposed teacher is therefore

$$q^{\text{recon}}_{T,t} := \text{softmax}_{\mathcal{V}}\big(z^0_{V,t} + \widehat{\alpha}_{S,t} u_t\big), \tag{51}$$

where the coefficient statistics use $q^0_{V,t}$ as the receiving prior. The direct-teacher condition uses $q^I_{V,t}$ as its target.

For the raw residual and exact calibration, this self-prior construction recovers the original image-conditioned teacher at $\rho = 1$ (Appendix A.4). The training comparison instead evaluates the complete practical recomposition, including coefficient approximation and stabilization. Its performance difference from standard OPD therefore reflects their combined effect.

Distillation results. Table 10 reports the complete reasoning results. The recomposed teacher obtains similar aggregate performance to direct VLM supervision (40.18 vs. 40.37 for multimodal reasoning and 51.18 vs. 51.72 for text reasoning), with mixed changes across benchmarks. This comparison supports the usefulness of the practical teacher for distillation, while not establishing equivalence or attributing individual gains to larger visual updates. Coefficient approximation and residual stabilization are evaluated together, so their separate training effects cannot be isolated by this comparison.

Table 10: **Practical recomposition: downstream distillation.** Both conditions train the same 2B student for 200 steps using the 4B VLM. Recomposition uses the VLM's own image-free prior, stabilized visual residual, and approximate calibration at $\rho = 1$. Scores are mean $\pm$ SD over five evaluation sampling seeds. Δ is recomposed minus direct-supervision accuracy in pp. Averages report means only.

Benchmark	**Direct VLM supervision**	**Recomposed teacher**	Δ (pp)
	Multimodal Reasoning		
MATH-Vision	27.43 ± 0.89	28.62 ± 1.73	$+1.19$
MathVerse (Total)	52.70 ± 0.76	53.43 ± 0.38	$+0.73$
MathVista	68.60 ± 0.39	68.34 ± 1.28	-0.26
MMMU-Pro	36.98 ± 0.89	36.72 ± 0.51	-0.26
RBench-M	16.90 ± 0.52	13.86 ± 1.27	-3.04
OlympiadBench-MM	39.60 ± 0.37	40.13 ± 2.72	$+0.53$
Multimodal Avg.	40.37	40.18	-0.19
	Text Reasoning		
MATH-500	78.72 ± 0.82	78.72 ± 1.85	$+0.00$
MathVerse (Text Only)	57.44 ± 0.82	56.80 ± 1.19	-0.64
MMLU-Pro	45.52 ± 0.16	45.62 ± 0.33	$+0.10$
RBench-T	28.59 ± 1.06	28.01 ± 1.05	-0.58
OlympiadBench-Text	48.31 ± 0.70	46.77 ± 1.98	-1.54
Text Avg.	51.72	51.18	-0.54

E SUPPLEMENTARY EXPERIMENTS

E.1 MAIN RESULTS AND DECODING VARIANTS

E.1.1 SAMPLED DECODING

Table 11 adds evaluation SDs to the reasoning results in Table 1. Scores are accuracies (%), reported using the five-seed evaluation protocol in Appendix C.3. For LEGO-OPD, this table reports $\rho = 1$; results at $\rho = 2$ with SDs are provided in Appendix D.2.

Table 11: **Main reasoning results with evaluation variability.** Scores are accuracies (%, mean $\pm$ SD across five sampling seeds). Base 2B is the initial student and Base 4B is the original VLM reference. All distilled models are 2B students evaluated at step 200. Averages report means only.

Benchmark	**Base 4B**	**Base 2B**	**Standard OPD**	**VA-OPD**	**M-OPD**	**H-OPD**	**LEGO-OPD** $\rho = 1$
			Multimodal Reasoning				
MATH-Vision	39.74 ± 2.25	27.96 ± 1.68	27.43 ± 0.89	25.33 ± 1.12	24.41 ± 1.73	32.37 ± 1.83	33.36 ± 2.56
MathVerse (Total)	66.70 ± 0.51	51.51 ± 0.48	52.70 ± 0.76	51.66 ± 0.36	51.28 ± 0.53	55.41 ± 0.65	56.96 ± 0.20
MathVista	76.26 ± 0.59	66.56 ± 0.71	68.60 ± 0.39	67.36 ± 0.93	66.42 ± 1.00	67.70 ± 0.79	69.62 ± 0.72
MMMU-Pro	50.29 ± 0.85	37.40 ± 0.66	36.98 ± 0.89	35.11 ± 1.16	35.06 ± 0.85	39.36 ± 0.69	40.98 ± 1.14
RBench-M	32.81 ± 1.34	19.61 ± 0.95	16.90 ± 0.52	16.81 ± 0.62	15.67 ± 1.30	26.92 ± 1.59	29.17 ± 1.20
OlympiadBench-MM	63.20 ± 2.02	37.60 ± 3.42	39.60 ± 0.37	37.33 ± 1.70	36.53 ± 1.66	39.07 ± 3.08	44.93 ± 2.24
Multimodal Avg.	54.83	40.11	40.37	38.93	38.23	43.47	45.84
			Text Reasoning				
MATH-500	93.04 ± 0.55	80.44 ± 0.95	78.72 ± 0.82	80.20 ± 0.69	80.56 ± 0.77	81.88 ± 1.08	82.40 ± 0.99
MathVerse (Text Only)	66.80 ± 0.17	57.18 ± 1.01	57.44 ± 0.82	56.65 ± 1.01	56.19 ± 0.79	59.44 ± 0.39	59.59 ± 0.99
MMLU-Pro	62.90 ± 0.14	43.59 ± 0.17	45.52 ± 0.16	45.09 ± 0.09	44.98 ± 0.30	46.94 ± 0.17	47.15 ± 0.05
RBench-T	51.92 ± 0.82	30.20 ± 0.31	28.59 ± 1.06	28.76 ± 0.52	28.28 ± 1.01	37.17 ± 1.12	38.59 ± 1.19
OlympiadBench-Text	69.41 ± 0.95	48.04 ± 1.42	48.31 ± 0.70	48.99 ± 0.73	47.77 ± 1.13	52.40 ± 0.95	52.31 ± 0.89
Text Avg.	68.81	51.89	51.72	51.94	51.56	55.57	56.01

E.1.2 GREEDY DECODING

Table 12 reports greedy-decoding results for the initial 2B student, the 4B VLM reference, and the distilled 2B students at step 200. All evaluation settings are identical to those used for sampling-based evaluation, except for the sampling parameters. Standard OPD denotes distillation from a single VLM. LEGO-OPD uses the 4B language and grounding experts at $\rho \in \{1, 2\}$. Under greedy decoding, LEGO-OPD at $\rho = 2$ achieves the highest multimodal and text reasoning averages among the evaluated 2B students.

Table 12: **Greedy decoding results across reasoning benchmarks.** Scores are accuracies (%). Base 2B and Base 4B are the original Qwen3-VL-Instruct models without distillation. All distilled students are evaluated at step 200. Averages are unweighted means over the six multimodal or five text benchmarks. Bold indicates the best score among the 2B students, excluding the 4B reference.

	4B reference	Initial student	Distillation baselines				LEGO-OPD	
Benchmark	**Base 4B**	**Base 2B**	**Standard OPD**	**VA-OPD**	**M-OPD**	**H-OPD**	$\rho = 1$	$\rho = 2$
Step	–	–	200	200	200	200	200	200
Multimodal Reasoning								
MATH-Vision	31.91	8.88	13.82	11.84	11.18	19.74	25.66	**28.95**
MathVerse (Total)	62.77	38.40	39.80	40.18	39.57	47.11	**53.78**	53.43
MathVista	75.10	58.70	60.20	60.70	62.50	65.30	68.30	**70.60**
MMMU-Pro	44.22	27.17	24.80	27.40	26.42	32.02	36.99	**37.34**
RBench-M	14.14	2.41	2.86	2.41	3.16	7.37	**18.35**	15.79
OlympiadBench-MM	64.00	26.00	32.00	29.33	28.67	**41.33**	40.00	40.00
Multimodal Avg.	48.69	26.93	28.91	28.64	28.58	35.48	40.51	**41.02**
Text Reasoning								
MATH-500	90.40	64.40	71.60	70.20	72.60	79.00	80.40	**82.00**
MathVerse (Text Only)	63.07	38.71	45.81	45.81	43.53	53.05	57.23	**58.50**
MMLU-Pro	38.75	43.26	24.72	38.22	38.11	36.69	44.82	**46.07**
RBench-T	33.27	6.58	9.14	10.42	9.41	20.66	28.43	**29.16**
OlympiadBench-Text	63.95	33.68	40.80	39.61	39.91	46.88	48.52	**48.66**
Text Avg.	57.89	37.33	38.41	40.85	40.71	47.26	51.88	**52.88**

E.2 Expert Configurations and Factor Contributions

Models and comparison settings. LLM 1.7B and LLM 4B denote Qwen3-1.7B and Qwen3-4B-Instruct-2507. VLM 2B/4B/8B denote the corresponding Qwen3-VL-Instruct models. All combinations train the same 2B student for 200 steps with $\rho = 1$ and the practical settings in Appendix B. A VLM used as the language expert receives no image. The grounding expert supplies both image-conditioned and image-free predictions for the visual residual. Table 13 provides the full reasoning results with evaluation SDs for Tables 4 and 3. Shared configurations are reported once. The 4B VLM self-recomposition condition is reported with the calibration analysis in Appendix D.6.

Model selection. The 1.7B language model and 2B/8B grounding models vary one expert while the other remains fixed at 4B. Checkpoint generation and inference mode also affect expert strength: Qwen3-4B-Instruct-2507 reports higher non-thinking AIME25 and ZebraLogic scores than the original Qwen3-8B (Yang et al., 2025; Qwen Team, 2025). Thus, a larger parameter count alone does not establish that the original 8B language model is a stronger replacement for the updated 4B checkpoint.

Expert-factor substitution. To examine the contributions of the two factor sources, we replace either the grounding expert or the language prior with the corresponding factor from the frozen initial student, as described in Section 4.3. Grounding-expert substitution uses the frozen student's image-conditioned and image-free predictions to construct the visual residual. Language-prior substitution instead uses the frozen student's image-free prediction as the receiving prior. The remaining expert is retained, and the calibration statistics are computed from the resulting factor sources.

Benchmark subset for the compact factor comparison. The compact comparison in Table 3 uses MATH-Vision, MathVerse, and OlympiadBench for multimodal evaluation. MATH-Vision represents visual mathematical reasoning and is also used in the baseline work. MathVerse provides fine-grained input variants, while OlympiadBench adds challenging olympiad-level reasoning. For text-only evaluation, we use the corresponding MathVerse and OlympiadBench benchmarks to compare reasoning across modalities within the same benchmark families.

Table 13: **Expert combinations with evaluation variability.** Column headers list the language expert above the grounding expert. The shaded Base column uses the 4B language and grounding experts. VLM 2B is the frozen initial student and receives no image when used as the language expert. Scores are mean ± SD over five evaluation sampling seeds. Averages report means only.

Benchmark	Language expert swap		Base	Grounding expert swap	
	LLM 1.7B VLM 4B	VLM 2B VLM 4B	LLM 4B VLM 4B	LLM 4B VLM 2B	LLM 4B VLM 8B
Multimodal Reasoning					
MATH-Vision	28.36 ± 2.84	26.58 ± 1.20	33.36 ± 2.56	33.03 ± 1.35	34.34 ± 1.77
MathVerse (Total)	43.89 ± 0.27	52.69 ± 0.54	56.96 ± 0.20	54.17 ± 0.65	57.81 ± 0.40
MathVista	64.96 ± 0.89	68.20 ± 0.47	69.62 ± 0.72	69.42 ± 1.13	70.24 ± 0.56
MMMU-Pro	35.11 ± 0.63	36.36 ± 0.77	40.98 ± 1.14	41.93 ± 0.69	41.70 ± 1.03
RBench-M	22.71 ± 1.48	17.53 ± 0.81	29.17 ± 1.20	28.48 ± 1.95	28.90 ± 1.40
OlympiadBench-MM	36.00 ± 2.21	37.20 ± 3.25	44.93 ± 2.24	43.07 ± 3.55	45.20 ± 1.28
Multimodal Avg.	38.51	39.76	45.84	45.02	46.37
Text Reasoning					
MATH-500	81.52 ± 0.89	79.60 ± 0.85	82.40 ± 0.99	82.36 ± 0.57	84.08 ± 1.81
MathVerse (Text Only)	55.03 ± 1.84	56.98 ± 0.68	59.59 ± 0.99	59.92 ± 0.61	59.59 ± 0.95
MMLU-Pro	43.14 ± 0.41	44.91 ± 0.18	47.15 ± 0.05	47.51 ± 0.35	46.92 ± 0.31
RBench-T	33.05 ± 1.36	29.87 ± 1.33	38.59 ± 1.19	38.46 ± 1.30	37.61 ± 0.90
OlympiadBench-Text	47.57 ± 1.53	48.07 ± 0.87	52.31 ± 0.89	52.67 ± 1.35	53.06 ± 0.99
Text Avg.	52.06	51.89	56.01	56.18	56.25

E.3 Decomposed OPD Reproduction Attempt

Implementation and results. At the time of our experiments, no official implementation of Decomposed OPD (Yoon et al., 2026) was available. We therefore implemented the method based on the paper. However, we were unable to reproduce its reported improvements in our experimental setting. We report this attempt separately from the main baseline comparisons.

We use Qwen3-VL-2B-Instruct as the student and Qwen3-VL-4B-Instruct as the teacher, with the training data and evaluation protocol described in Section 4.1. The teacher is evaluated with and without the image. Our implementation includes the standard distillation, visual-steering, and language-preservation terms, with $\gamma = 2.0$, $\lambda = 0.01$, and $\eta = 0.41$. The language-preservation mask uses the per-response 70th percentile of the teacher visual-dependence score. All distillation terms are computed over the full vocabulary, whereas our Standard OPD baseline uses the student-top-16 surrogate.

Table 14 reports the results. Our implementation obtains lower average scores than Standard OPD on both the multimodal and text-only benchmarks. These results describe our implementation under the adapted setup and do not establish the performance of the original method under its original experimental conditions.

Table 14: Results of our Decomposed OPD reproduction attempt at step 200. Scores are accuracies (%). Reasoning scores report mean ± SD. Perception scores use greedy decoding. Benchmark averages report means only, computed as unweighted means over six multimodal or five text benchmarks. Decomposed OPD denotes our paper-based implementation.

Benchmark	Standard OPD	Decomposed OPD (our implementation)
Multimodal reasoning		
MATH-Vision	27.43 ± 0.89	25.59 ± 1.82
MathVerse (Total)	52.70 ± 0.76	50.80 ± 0.30
MathVista	68.60 ± 0.39	63.94 ± 1.28
MMMU-Pro	36.98 ± 0.89	34.52 ± 0.56
RBench-M	16.90 ± 0.52	10.20 ± 0.91
OlympiadBench (MM)	39.60 ± 0.37	35.60 ± 1.74
Multimodal reasoning Avg.	40.37	36.78
Text-only reasoning		
MATH-500	78.72 ± 0.82	79.36 ± 1.23
MathVerse (Text Only)	57.44 ± 0.82	56.24 ± 1.23
MMLU-Pro	45.52 ± 0.16	43.64 ± 0.23
RBench-T	28.59 ± 1.06	22.23 ± 1.07
OlympiadBench (Text)	48.31 ± 0.70	46.56 ± 0.68
Text-only reasoning Avg.	51.72	49.61
Visual perception		
BLINK	52.13	51.60
QBench1,2	69.54	69.18

The relation between its auxiliary visual objective and LEGO-OPD's loss decomposition is derived in Appendix A.5.

F LIMITATIONS

Although LEGO-OPD calibrates how strongly visual evidence modifies the language prior, it does not determine whether that evidence is reliable. Source KL measures the magnitude of the grounding expert's visual update rather than its correctness, so calibration can amplify erroneous evidence.

Beyond calibration, the visual residual represents a conditional model contrast rather than an isolated visual factor, because generated prefixes may already contain information inferred from the image. In addition, the current implementation requires a shared tokenization and output vocabulary, restricting composition to compatible models within the same family.