LEARN IN SPACE: ON-POLICY DISTILLATION VIA SPATIAL DIVERGENCE FOR VISION-CENTRIC REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial awareness matters in vision-centric reasoning. Despite the effectiveness of its fine-grained supervision, on-policy distillation (OPD) remains spatially agnostic during optimization. Fundamentally, the same visual evidence can admit many spatially equivalent region descriptions, consequently leading to a vast number of distinct token serializations. Token-level KL treats these serializations as distinct outcomes in vocabulary space, which provides biased learning signals for spatially valid grounding behavior, ultimately limiting performance. To bridge these gaps, we introduce MixOPD, which uses *Spatial Divergence* to measure the geometric discrepancy between each on-policy student grounding proposal and corresponding teacher-supported evidence distribution estimated via *Teacher-Forced Constrained Search*, thereby spatially calibrating the token-level KL objective. MixOPD integrates seamlessly into typical OPD frameworks and delivers theoretically grounded performance gains. Extensive experiments demonstrate that MixOPD consistently improves both standard and entropy-aware OPD. It achieves highly competitive performance against representative post-training baselines and recent alternatives, while presenting favorable inference latency and stable training dynamics.

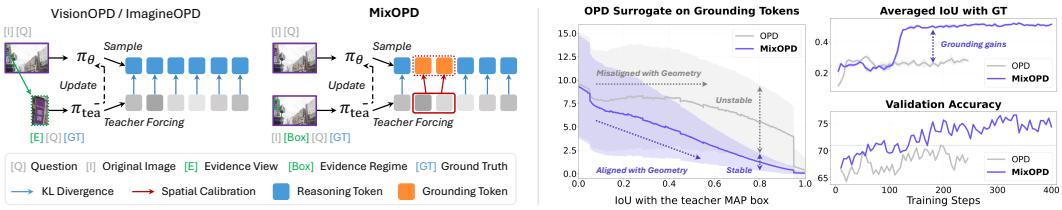

Figure 1: MixOPD brings explicit spatial credit assignment to OPD-based vision-centric reasoning. *Left:* Existing methods distill evidence-aware reasoning from a teacher π_{tea} conditioned on a zoomed-in view into the student policy π_θ via token-level divergence, whereas MixOPD interleaves explicit grounding with reasoning and spatially calibrates the local surrogate at coordinate-bearing positions. *Right:* Unlike standard OPD, spatial calibration aligns the grounding-token surrogate with geometry: as student–teacher spatial agreement improves, its mean decreases smoothly and monotonically while its variance contracts, translating into higher grounding precision and validation accuracy.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) achieve remarkable performance across diverse visual tasks Bai et al. (2025a;b); Liu et al. (2024); however, they still face challenges with interpretability and hallucination, particularly in complex scenarios, *e.g.*, fine-grained visual understanding. To address these challenges, prior works Wang et al. (2025a;b); Liu et al. (2025); Li et al. (2026a) rely on reinforcement learning, *e.g.*, GRPO Guo et al. (2025), to train LVLMs to predict evidence coordinates as visual anchors for grounding intermediate thoughts. However, most of these methods suffer from sparse sequence-level rewards, restricting data utilization and optimization efficiency.

Recent advances, *e.g.*, VisionOPD Yuan et al. (2026) and ImagineOPD Cai et al. (2026), adopt On-Policy Distillation (OPD) Lu & Lab (2025) to provide fine-grained supervision for visual reasoning in LVLMs. They use the detached reverse-KL signal from a fixed privileged teacher π_{tea} as a *token-level*

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.

advantage within a PPO-clipped surrogate, efficiently transferring evidence-aware reasoning into the student policy π_θ, as shown in Figure 1 (*left*). Despite these advances, they leave evidence selection implicit: the student may improve the distributional agreement by exploiting shortcuts or distractors rather than learning the intended grounding behavior, leading to weak visual credit assignment.

A natural direction is to interleave explicit grounding with reasoning, thereby extending the dense supervision provided by OPD to both. However, this strategy faces a fundamental mismatch: standard OPD provides *token-wise* supervision at coordinate-bearing positions, whereas grounding quality is determined jointly by *span-level* geometry. Intuitively, the same visual evidence may admit many spatially equivalent yet numerically distinct realizations. When the teacher spreads probability mass across these valid realizations, OPD's token-level advantages become serialization-sensitive rather than geometry-aligned: a geometrically better proposal need not receive a higher aggregate advantage, while proposals of comparable geometric quality can receive markedly different advantages.

To address this mismatch, we present MixOPD, a spatially aware OPD framework built on *span-to-token* calibration. Specifically, for each well-formed grounding proposal, MixOPD defines a *Spatial Divergence* that measures the geometric discrepancy between the proposal and its corresponding teacher-supported box distribution, estimated online via *teacher-forced constrained coordinate search*. The resulting span-level coefficient is then broadcast across all coordinate-bearing positions in the span to reweight the full local surrogate. Empirically, this calibration aligns the grounding-token surrogate with proposal geometry, yielding higher grounding precision and validation accuracy, as illustrated in Figure 1 (*right*). The same construction applies to different OPD formulations, including standard OPD and entropy-aware OPD (EOPD), and admits a direct theoretical guarantee (Them. 3.8).

By introducing spatial calibration into the joint optimization of grounding and reasoning, MixOPD achieves state-of-the-art performance across diverse visual tasks. On fine-grained benchmarks such as V* Bench Wu & Xie (2024) and HR-Bench Wang et al. (2025c), it outperforms the base models by 3.6–14.9 percentage points at both scales (4B / 8B), and further demonstrates strong out-of-distribution generalization to general-purpose VQA benchmarks, including MME-RealWorld-CN/EN Zhang et al. (2025b), MMStar Chen et al. (2024), and CV-Bench Tong et al. (2024). These gains remain consistent and substantial under both standard OPD Lu & Lab (2025) and EOPD Jin et al. (2026) formulations and are accompanied by more stable training dynamics.

2 Background

2.1 Related Work

Vision-centric reasoning (*a.k.a.* o1-style thinking with images) seeks to improve the capabilities and reliability of LVLMs in VQA by incorporating explicit evidence localization as visual anchors into the reasoning process. Specifically, DeepEyes series Zheng et al. (2025); Hong et al. (2025a) and Thyme Zhang et al. (2025a) explore evidence localization through sandboxed tool use and code execution, respectively. Another line of work Wang et al. (2025a); Liu et al. (2025); Shen et al. (2025); Li et al. (2026a); Luo et al. (2025) treats self-generated evidence coordinates as intermediate visual thoughts and optimizes policies by enforcing geometric consistency between predicted coordinates and spatial priors. More recent studies Lu et al. (2026); Zhang et al. (2026a) further establish visual grounding as a primitive for vision-centric reasoning. Despite these advances, existing methods largely rely on sparse rewards that separately measure grounding precision and reasoning correctness. Such a paradigm provides limited credit for whether a particular grounding behavior actually supports the following reasoning. To address this, MixOPD provides interdependent token-level supervision in both grounding and reasoning positions, thereby delivering more effective joint optimization.

On-policy distillation (OPD) Agarwal et al. (2024); Lu & Lab (2025) has emerged as an effective post-training paradigm for LLMs, combining on-policy rollouts with dense token-level supervision. Prior work advances OPD along several dimensions. REOPOLD Ko et al. (2026), BRTS Zhang et al. (2026b), and EDGE-OPD Lazaridis et al. (2026) refine supervision through relaxed imitation, best-of-N selection, and evidence-guided masking, respectively. OPSD Zhao et al. (2026) removes the dependency on an external teacher through privileged-context self-distillation, while EOPD Jin et al. (2026) introduces entropy-aware divergence selection for more robust optimization at token positions with multimodal supervision. VisionOPD Yuan et al. (2026) and ImagineOPD Cai et al. (2026) instantiate OPSD in the visual domain, which treats zoomed-in evidence as privileged context and internalizes evidence-aware reasoning. These methods improve how visual evidence informs

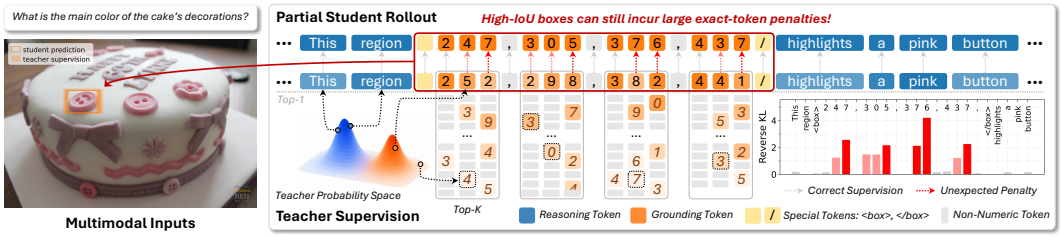

Figure 2: Illustration of token-level KL limitation in vision-centric reasoning. The teacher distributes probability mass across spatially neighboring serializations rather than concentrating it on the student's exact grounding span, imposing spurious per-token penalties on spatially valid grounding behaviors.

reasoning, but provide no explicit supervision for where the model should look. MixOPD closes this gap with spatially calibrated supervision over explicit evidence localizations, which enables LVLMs to actively ground task-relevant evidence, inducing better reasoning performance.

2.2 Preliminary

On-policy distillation (OPD) Agarwal et al. (2024) trains a student on its own rollouts using dense teacher signals. For $\mathbf{q} \sim \mathcal{Q}$, set $\pi_{\theta_{\text{old}}} = \text{sg}(\pi_\theta)$, sample $\mathbf{x} = (x_1, \ldots, x_T) \sim \pi_{\theta_{\text{old}}}(\cdot \mid \mathbf{q})$, and define the student context $\mathbf{c}_t = (\mathbf{q}, \mathbf{x}_{<t})$ and privileged teacher context $\widetilde{\mathbf{c}}_t$. Standard OPD Lu & Lab (2025) uses the detached sampled reverse-KL signal as a token-level advantage:

$$\widehat{A}_t = \text{sg}\left[\text{clip}\left(\log \frac{\pi_{\text{tea}}(x_t \mid \widetilde{\mathbf{c}}_t)}{\pi_{\theta_{\text{old}}}(x_t \mid \mathbf{c}_t)}, -C_A, C_A\right)\right], \qquad \varrho_t(\theta) = \frac{\pi_\theta(x_t \mid \mathbf{c}_t)}{\pi_{\theta_{\text{old}}}(x_t \mid \mathbf{c}_t)}.$$

For a minibatch $\mathcal{M}$ with token mask m_t, it minimizes the globally token-averaged PPO surrogate

$$\mathcal{L}_{\text{OPD}}(\theta) = \mathbb{E}_{\mathcal{M}}\left[\frac{\sum_{\mathcal{M}} m_t \ell_t^{\text{OPD}}(\theta)}{\sum_{\mathcal{M}} m_t}\right], \quad \ell_t^{\text{OPD}} = \max\left\{-\varrho_t \widehat{A}_t, -\text{clip}(\varrho_t, 1\text{-}\epsilon_{\text{clip}}, 1\text{+}\epsilon_{\text{clip}})\widehat{A}_t\right\}. \quad (1)$$

Here $C_A > 0$ and $\epsilon_{\text{clip}} \in (0, 1)$ are the sampled-log-ratio and PPO clipping radii. At $\theta = \theta_{\text{old}}$, when the former is inactive on the rollout-policy support, the expected semigradient matches reverse KL.

Definition 2.1 (Reverse and Forward KL Divergence). *For $p_t = \pi_\theta(\cdot \mid \mathbf{c}_t)$ and $q_t = \pi_{\text{tea}}(\cdot \mid \widetilde{\mathbf{c}}_t)$,*

$$\mathcal{D}_{\text{RKL}}(p_t, q_t) = \mathbb{E}_{x \sim p_t}\left[\log \frac{p_t(x)}{q_t(x)}\right], \qquad \mathcal{D}_{\text{FKL}}(q_t, p_t) = \mathbb{E}_{x \sim q_t}\left[\log \frac{q_t(x)}{p_t(x)}\right].$$

Remark 2.2. *The two KL directions induce distinct learning dynamics. Reverse KL penalizes student probability mass assigned to low-probability regions of the teacher and therefore favors its dominant modes; forward KL penalizes missing teacher-supported modes and encourages broader coverage.*

EOPD Jin et al. (2026) retains ℓ_t^{OPD} and adds forward KL at high-entropy teacher positions. With $g_t = \mathbf{1}[H(q_t) > \eta_H]$ and $\bar{q}_t^L$ denoting q_t renormalized over its Top-L support, define $S_t^{(\lambda_{\text{F}})}(\theta) = \ell_t^{\text{OPD}}(\theta) + \lambda_{\text{F}} g_t D_{\text{KL}}(\bar{q}_t^L \| p_t)$. Here, $\lambda_{\text{F}} = 0$ recovers standard OPD, whereas $\lambda_{\text{F}} > 0$ yields EOPD.

3 Methodology

3.1 Limitations of KL Divergence in Vision-Centric Reasoning

Standard OPD is misaligned with visual grounding in two ways. First, it is *not spatially aware*: similar token discrepancies can encode markedly different localization errors, while distinct serializations can decode to neighboring regions. Second, one evidence region may admit multiple valid coordinate realizations, yet token-level matching also constrains how probability is allocated among them (Fig. 2). We formalize this mismatch on the coordinate decisions that MixOPD actually calibrates.

Definition 3.1 (Grounding Span and Coordinate-Decision Law). *Consider a trajectory containing J well-formed grounding spans $\{\tau(b_j)\}_{j=1}^J$, where a_j is the start of span j and $\mathcal{I}(b_j)$ its coordinate-bearing positions. Fix a finite deterministic box grammar and a finite quantized box space $\mathcal{B}$. The grammar inserts structural tokens and exposes only coordinate-bearing decision nodes. Each canonical box $b \in \mathcal{B}$ has a distinct stopped decision trace $\iota(b)$; all other traces are malformed. At each*

decision node, policy π_u retains its full-vocabulary distribution on the grammar-completed, candidate-specific prefix. This defines a normalized raw trace law $\mathsf{P}_{u,\mathcal{I}}(y \mid \mathbf{c}) = \prod_{\ell=1}^{L_{\mathcal{I}}} K^{\mathbf{c}}_{u,\ell}(y_\ell \mid y_{<\ell})$, where $K^{\mathbf{c}}_{u,\ell}$ is the induced full-vocabulary kernel and $L_{\mathcal{I}}$ the padded stopping depth, with an absorbing stop state. A parser maps each canonical trace to $\iota(b)$ and every noncanonical trace to a malformed atom $\perp$. Its normalized pushforward $P^{\mathcal{I}}_u \in \mathcal{P}(\Omega_{\mathbf{c}})$, where $\Omega_{\mathbf{c}} = \iota(\mathcal{B}) \cup \{\perp\}$, satisfies

$$P^{\mathcal{I}}_u(\iota(b) \mid \mathbf{c}) = \prod_{\ell=1}^{|\iota(b)|} K^{\mathbf{c}}_{u,\ell}(\iota(b)_\ell \mid \iota(b)_{<\ell}), \qquad P^{\mathcal{I}}_u(\perp \mid \mathbf{c}) = 1 - \textstyle\sum_{b\in\mathcal{B}} P^{\mathcal{I}}_u(\iota(b) \mid \mathbf{c}).$$

For a realized b_j, this atom's log-density ratio is exactly the sum over its coordinate-bearing positions $\mathcal{I}(b_j)$. The construction intervenes on, rather than conditions upon, structural tokens; their probabilities are absent from both this law and MixOPD's calibration.

Definition 3.2 (Contextual Evidence Quotient)**.** *Fix $\mathbf{c}$, let $e(\mathbf{c})$ be the targeted visual evidence, assume $\mathcal{B}_{\mathrm{val}}(\mathbf{c}) \neq \varnothing$, and partition $\mathcal{B}_{\mathrm{val}}(\mathbf{c}) = \{b \in \mathcal{B} : b \text{ visually grounds } e(\mathbf{c})\}$, and $\mathcal{B}_{\mathrm{inv}}(\mathbf{c}) = \mathcal{B}\backslash\mathcal{B}_{\mathrm{val}}(\mathbf{c})$. The valid set has Jaccard diameter $\varepsilon(\mathbf{c}) = \sup_{b,b'\in\mathcal{B}_{\mathrm{val}}(\mathbf{c})}[1 - \mathrm{IoU}(b, b')]$. Define $\kappa_{\mathbf{c}} : \Omega_{\mathbf{c}} \to \mathcal{Z}(\mathbf{c}) = \{e(\mathbf{c}), \perp\}$ by mapping $\iota(b)$ to $e(\mathbf{c})$ iff $b \in \mathcal{B}_{\mathrm{val}}(\mathbf{c})$, and mapping invalid traces and $\perp$ to $\perp$. Its normalized pushforward is $P^{\mathcal{Z}}_u(e \mid \mathbf{c}) = \sum_{b\in\mathcal{B}_{\mathrm{val}}(\mathbf{c})} P^{\mathcal{I}}_u(\iota(b) \mid \mathbf{c})$, and $P^{\mathcal{Z}}_u(\perp \mid \mathbf{c}) = 1 - P^{\mathcal{Z}}_u(e \mid \mathbf{c})$.*

This separates grounding validity from allocation among exact coordinate realizations. Applying the KL chain rule to the partition induced by $\kappa_{\mathbf{c}}$ gives the following.

Proposition 3.3 (Evidence-Quotient Decomposition)**.** *For $z \in \mathcal{Z}(\mathbf{c})$ with $P^{\mathcal{Z}}_u(z \mid \mathbf{c}) > 0$, define $P^{\mathcal{I}}_u(s \mid z, \mathbf{c}) = P^{\mathcal{I}}_u(s \mid \mathbf{c})/P^{\mathcal{Z}}_u(z \mid \mathbf{c})$, where $\kappa_{\mathbf{c}}(s) = z$. For $P^{\mathcal{I}}_u \ll P^{\mathcal{I}}_v$, we suppress $\mathbf{c}$ and have*

$$D_{\mathrm{KL}}(P^{\mathcal{I}}_u \| P^{\mathcal{I}}_v) = D_{\mathrm{KL}}(P^{\mathcal{Z}}_u \| P^{\mathcal{Z}}_v) + \mathbb{E}_{z\sim P^{\mathcal{Z}}_u}\left[D_{\mathrm{KL}}\left(P^{\mathcal{I}}_u(\cdot \mid z) \middle\| P^{\mathcal{I}}_v(\cdot \mid z)\right)\right].$$

Remark 3.4. *The first term measures disagreement in grounding validity, while the second captures residual disagreement among exact coordinate realizations within each class. This reveals that coordinate-level matching used in standard OPD can even penalize spatially equivalent behavior.*

Proof. This result follows from Prop. E.8; supporting empirical evidence is presented in §A.1. □

3.2 Calibration with Spatial Divergence

To address the limitation (§3.1), MixOPD first formulates a *spatial divergence* that quantifies the span-level geometric discrepancy between each realized proposal and its corresponding supervision, then practically estimates it to calibrate the local OPD surrogate for spatially aware optimization.

Divergence Formulation. For a realized grounding span $\tau(b_j)$ in a student rollout, let $B^f_j \in \mathcal{P}(\mathcal{B})$ denote its spatial supervision law. We introduce *Spatial Divergence*, which measures the discrepancy between b_j and B^f_j as an expected *point-to-distribution* risk: $\mathcal{D}_{\mathrm{spa}}(\delta_{b_j}, B^f_j) := \mathbb{E}_{\widetilde{b}\sim B^f_j}\left[1 - \mathrm{IoU}(b_j, \widetilde{b})\right]$.

Even with sample-level evidence annotations, a span-aligned ground-truth law B^f_j is generally unavailable: an on-policy span may target indirect evidence outside the annotated set, while multiple relevant annotations may leave its one-to-one correspondence ambiguous. Fortunately, OPD provides a natural surrogate: the *ideal* teacher box law $B^{\mathrm{exp}}_j := P^{\mathcal{B}}_{\mathrm{tea}}(\cdot \mid \widetilde{\mathbf{c}}_{a_j})$, obtained by normalizing the well-formed atoms of $P^{\mathcal{I}}_{\mathrm{tea}}(\cdot \mid \widetilde{\mathbf{c}}_{a_j})$. Two direct alternatives either collapse this law to a MAP point mass, $B^{\mathrm{MAP}}_j := \delta_{b^{\mathrm{MAP}}_j}$, where $b^{\mathrm{MAP}}_j \in \arg\max_b B^{\mathrm{exp}}_j(b)$, or flatten it over its support, $B^{\mathrm{uni}}_j := U_{\mathrm{supp}(B^{\mathrm{exp}}_j)}$. Empirically, *teacher probabilities exhibit measurable geometric selectivity, and retaining them yields more robust calibration*, as detailed in §A.2. We therefore adopt $B^f_j = B^{\mathrm{exp}}_j$ in this work.

Practical Approximation. Since exact enumeration of $P^{\mathcal{B}}_{\mathrm{tea}}$ is generally intractable, a direct autoregressive approximation is constrained *beam search*. However, it requires expensive candidate-conditioned sequential decoding for every grounding span, making it impractical for online training. To overcome this, we present *teacher-forced constrained search*, which reuses cached position-wise scores from the rollout-level teacher-forcing pass for bounded coordinate composition. From this single pass, it retains only digit tokens from each position's Top-V list, keeps the Top-M_{c} partial values after every digit extension, and combines them through axis and Cartesian merges with cutoff K under the realized span template and geometric constraints (detailed in §B.2). Empirically, it closely approximates the spatial divergence obtained via candidate-specific autoregressive rescoring on the same retained support and exhibits substantial end-to-end agreement with width-128 constrained beam search, while reducing the measured span-wise search overhead by over three orders

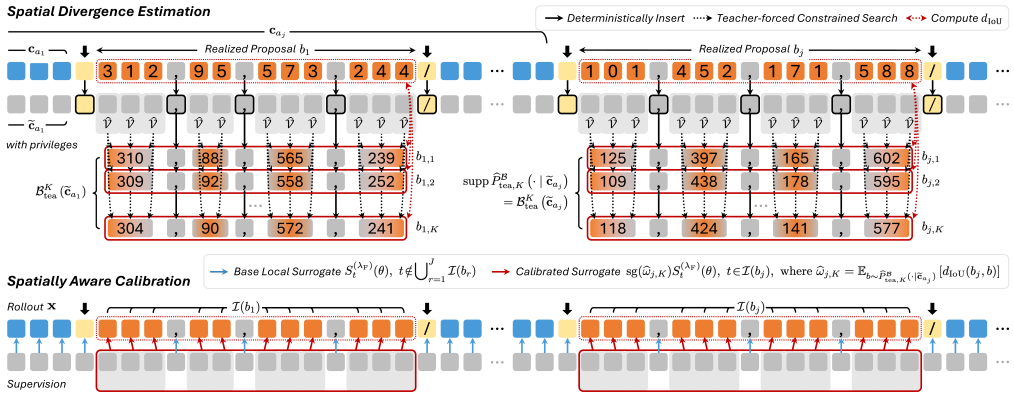

Figure 3: MixOPD estimates the spatial supervision for each proposal via a *teacher-forced constrained search*, then calibrates the implemented local surrogate $S_t^{(\lambda_{\mathrm{F}})}$ at coordinate-bearing positions via their *spatial divergence*. Here, $\widehat{\mathcal{V}} \subseteq \mathcal{V}$ denotes the set of tokens corresponding to the Top-V logits.

of magnitude (§A.3). Successful search retains at most K admissible boxes $\mathcal{B}^K_{\mathrm{tea}}(\widetilde{\mathbf{c}}_{a_j})$ with factorized scores $u_j(b)$. Normalizing these scores over the retained boxes yields the practical teacher box law

$$\widehat{P}^{\mathcal{B}}_{\mathrm{tea},K}(b \mid \widetilde{\mathbf{c}}_{a_j}) = \exp u_j(b) / \textstyle\sum_{\widetilde{b} \in \mathcal{B}^K_{\mathrm{tea}}(\widetilde{\mathbf{c}}_{a_j})} \exp u_j(\widetilde{b}), \qquad b \in \mathcal{B}^K_{\mathrm{tea}}(\widetilde{\mathbf{c}}_{a_j}), \tag{2}$$

and practical spatial coefficient $\widehat{\omega}_{j,K} = \mathcal{D}_{\mathrm{spa}}\big(\delta_{b_j}, \widehat{P}^{\mathcal{B}}_{\mathrm{tea},K}(\cdot \mid \widetilde{\mathbf{c}}_{a_j})\big)$. If search failed, we set $\widehat{\omega}_{j,K} = 1$.

Spatial Calibration. Let $\mathcal{J}_{\mathrm{suc}} \subseteq \{1, \ldots, J\}$ index the well-formed student grounding spans for which the practical search succeeds, and let $\{\mathcal{I}(b_j)\}_{j \in \mathcal{J}_{\mathrm{suc}}}$ be their pairwise-disjoint coordinate-bearing position sets. MixOPD applies one detached span coefficient to each such position:

$$\mathcal{L}^{(\lambda_{\mathrm{F}})}_{\mathrm{MixOPD}}(\theta) = \mathbb{E}_{\mathcal{M}}\left[\frac{\sum_{\mathcal{M}} m_t w_t S_t^{(\lambda_{\mathrm{F}})}(\theta)}{\sum_{\mathcal{M}} m_t}\right], \; w_t = \begin{cases} \mathrm{sg}(\widehat{\omega}_{j,K}), & \exists_{=1}\, j \in \mathcal{J}_{\mathrm{suc}} : t \in \mathcal{I}(b_j), \\ 1, & \text{otherwise.} \end{cases} \tag{3}$$

Therefore, for the reasoning and structural positions, as well as positions belonging to malformed, unparsed, or search-failed spans, retain unit weight. At each successfully searched, well-formed grounding span, MixOPD rescales the base local gradient by $\widehat{\omega}_{j,K} \in [0, 1]$, preserving its direction whenever $\widehat{\omega}_{j,K} > 0$ and suppressing it entirely only when $\widehat{\omega}_{j,K} = 0$.

3.3 Theoretical Analysis

We separate a distributional diagnosis from an implementation-faithful guarantee. Lemma 3.6 quantifies exact-coordinate pressure before practical clipping; Theorem 3.8 proves geometry-selective local semigradient scaling for the actual surrogate in Eq. 3. Fix $\mathbf{c} = \mathbf{c}_{a_j}$.

Assumption 3.5 (Bounded Valid-Realization Dilution)**.** *The teacher assigns total mass $\rho \geq \rho_{\min} > 0$ to $M \geq 2$ distinct valid coordinate realizations $\{b_{j,m}\}_{m=1}^{M} \subseteq \mathcal{B}_{\mathrm{val}}(\mathbf{c})$. Writing $q_m = P^{\mathcal{I}}_{\mathrm{tea}}(\iota(b_{j,m}) \mid \widetilde{\mathbf{c}})$ and $\sum_m q_m = \rho$, for fixed $\delta \in (0, 1)$ and $1 \leq \Gamma \leq M^{1-\delta}$, the following holds*

$$\tfrac{\rho}{\Gamma M} \leq q_m \leq \tfrac{\Gamma \rho}{M}, \; m = 1, \ldots, M.$$

This permits controlled non-uniformity without concentration on only a few realizations. Figure 5, 16 and 17 provide empirical entropy signatures consistent with this branching structure.

Lemma 3.6 (Dilution-Induced Coordinate Pressure)**.** *Under Assumption 3.5, let a valid student proposal be $b_j = b_{j,m^\star}$ and $p_j = P^{\mathcal{I}}_{\theta_{\mathrm{old}}}(\iota(b_j) \mid \mathbf{c}) \geq p_{\min} > 0$, where $p_{\min}$ is independent of M. Its unclipped realized coordinate log-density ratio $r_j^{\mathrm{coord}} = \log \frac{p_j}{q_{m^\star}} = \sum_{t \in \mathcal{I}(b_j)} \log \frac{\pi_{\theta_{\mathrm{old}}}(x_t \mid \mathbf{c}_t)}{\pi_{\mathrm{tea}}(x_t \mid \widetilde{\mathbf{c}}_t)}$ satisfies*

$$\delta \log M + \log p_{\min} \leq r_j^{\mathrm{coord}} \leq (2 - \delta) \log M + \log \tfrac{1}{\rho_{\min}}.$$

Remark 3.7. *These are finite-M bounds. Along any problem family with fixed $\delta, p_{\min}, \rho_{\min}$ and increasing M, the pressure is $\Theta(\log M)$. At $\theta = \theta_{\mathrm{old}}$, if neither sampled-log-ratio nor PPO clipping is active, $\sum_{t \in \mathcal{I}(b_j)} \ell_t^{\mathrm{OPD}} = r_j^{\mathrm{coord}}$. With clipping, r_j^{coord} remains a diagnostic of the underlying coordinate signal. Based on this, Theorem 3.8 directly characterizes the implemented surrogate.*

Table 1: Comparison with competitive alternatives. We report accuracy (%) for each model. The best results within each comparable group are highlighted in **bold**, and the second-best are underlined. VisionOPD† is *reproduced* with *Qwen3-VL 4B/8B* following the official recipe for fair comparisons.

Model	#Param	Fine-Grained Visual Tasks				General-Purpose Visual Tasks			
		VStar	Zoom	HR_{4K}	HR_{8K}	MME^{rw}_{CN}	MME^{rw}_{EN}	MMStar	CV-B
Agentic Framework for "Thinking with Images"									
DeepEyes Zheng et al. (2025)	7B	85.17	46.63	74.92	72.46	63.35	63.62	64.42	76.08
DeepEyesV2 Hong et al. (2025a)	7B	80.98	45.09	77.42	73.50	65.03	65.39	69.58	76.23
Thyme Zhang et al. (2025a)	7B	82.90	44.34	77.79	72.25	65.55	64.28	66.15	78.38
PFlowNet Li et al. (2026a)	8B	90.23	**48.72**	80.25	77.04	66.37	68.52	**72.05**	**87.88**
SenseNova MARS Chng et al. (2025)	8B	**92.84**	47.45	**83.92**	**78.04**	**67.58**	**69.22**	71.16	87.35
Private Large Vision-Language Models									
GPT-5.2 OpenAI (2025)	-	79.75	50.53	81.00	78.04	73.36	68.98	74.16	84.93
GPT-5.4 OpenAI (2026)	-	76.27	52.94	84.79	77.63	74.49	71.12	74.98	84.47
Gemini-3.1 Pro DeepMind (2026a)	-	88.31	60.43	**89.42**	86.13	**77.28**	73.53	83.05	**89.78**
Gemini-3.5 Flash DeepMind (2026b)	-	**89.35**	**60.67**	89.25	**86.96**	75.64	**74.19**	**83.75**	89.55
Open-Source Large Vision-Language Models									
GLM-4.6V Hong et al. (2025b)	106B	86.91	49.78	82.34	79.04	65.90	65.81	76.91	87.89
Qwen3-VL Bai et al. (2025a)	235B	**91.45**	**56.37**	**86.25**	**80.13**	**71.41**	**69.27**	76.15	86.96
Kimi-K2.6 Kimi Team (2026)	1T	88.13	53.41	81.67	78.17	68.47	65.81	**81.49**	**89.47**
Qwen3-VL Bai et al. (2025a)	4B	80.11	40.51	77.79	72.63	63.61	63.94	**70.44**	85.18
ZwZ Wei et al. (2026)	4B	87.78	54.08	80.79	**80.00**	65.58	66.00	**70.44**	**87.86**
VisionOPD† Yuan et al. (2026)	4B	82.90	43.08	78.00	74.75	65.97	**68.95**	69.82	85.76
ImagineOPD Cai et al. (2026)	4B	82.72	44.38	79.29	79.50	64.75	66.65	66.09	86.22
MixOPD (Ours)	4B	**89.70**	**55.38**	**81.42**	79.38	**66.18**	68.67	**70.42**	**87.91**
MiMo-VL RL Yue et al. (2025)	7B	82.90	44.93	73.04	70.13	62.41	55.16	**73.31**	84.43
MiniCPM-V 4.5 Yu et al. (2026b)	9B	71.03	42.48	70.09	60.75	62.98	61.41	68.18	79.97
Qwen3-VL Bai et al. (2025a)	8B	84.47	43.32	80.42	75.42	63.52	64.83	71.11	85.17
ZwZ Wei et al. (2026)	8B	87.26	56.96	83.84	**82.50**	66.28	**68.82**	71.85	**87.89**
VisionOPD† Yuan et al. (2026)	8B	84.64	46.39	79.21	75.96	64.18	65.74	71.49	85.27
ImagineOPD Cai et al. (2026)	8B	83.07	47.22	77.92	76.79	65.55	67.62	66.04	86.29
MixOPD (Ours)	8B	**90.40**	**57.83**	**84.50**	80.92	**66.53**	68.27	72.35	87.39

Let $d_{\mathrm{val}}(b;\mathbf{c}) = \inf_{b'\in\mathcal{B}_{\mathrm{val}}(\mathbf{c})}[1-\mathrm{IoU}(b,b')]$ and, for the clean statement, assume $\operatorname{supp}\widehat{P}^{\mathcal{B}}_{\mathrm{tea},K} \subseteq \mathcal{B}_{\mathrm{val}}(\mathbf{c})$. The finite-search and contaminated-support cases are treated in Lemma E.31 and E.33. Define the actual base and calibrated coordinate-span surrogates

$$S_j^{(\lambda_F)}(\theta) = \textstyle\sum_{t\in\mathcal{I}(b_j)} S_t^{(\lambda_F)}(\theta), \qquad \widetilde{S}_j^{(\lambda_F)}(\theta) = \mathrm{sg}(\widehat{\omega}_{j,K}) S_j^{(\lambda_F)}(\theta).$$

Theorem 3.8 (Geometry-Selective Local Semigradient Scaling)**.** *Under the valid-support condition above, fix the rollout and all detached teacher- and search-side quantities. For any local (sub)gradient $G_j \in \partial_\theta S_j^{(\lambda_F)}$ selected by the implemented surrogate, its calibrated counterpart satisfies*

$$\widetilde{G}_j = \widehat{\omega}_{j,K} G_j, \qquad \|\widetilde{G}_j\| = \widehat{\omega}_{j,K}\|G_j\|.$$

Let $\varepsilon = \varepsilon(\mathbf{c})$. Then:

(i) Valid-proposal attenuation. *If $b_j \in \mathcal{B}_{\mathrm{val}}(\mathbf{c})$, then $0 \le \widehat{\omega}_{j,K} \le \varepsilon$ and $\|\widetilde{G}_j\| \le \varepsilon\|G_j\|$. Thus calibration retains at most an ε-fraction of the local semigradient magnitude on a spatially valid proposal, suppressing at least a $1-\varepsilon$ fraction of the potentially spurious exact-token pressure.*

(ii) Separated-error retention. *For $\Delta\in[0,1]$, if $b_j\in\mathcal{B}_{\mathrm{inv}}(\mathbf{c})$ and $d_{\mathrm{val}}(b_j;\mathbf{c})\ge\Delta$, then $\Delta\le\widehat{\omega}_{j,K}\le 1$ and $\|\widetilde{G}_j\| \ge \Delta\|G_j\|$. These identities and bounds hold componentwise for the persistent PPO-OPD and additive forward-KL terms. Thus, if $\Delta > \varepsilon$, the scaling factor of every Δ-separated error exceeds that of any valid proposal by at least $\Delta-\varepsilon$.*

Proofs. Lem 3.6 follows from Lem E.15 (*i.e.*, $\beta=0, p_0=p_{\min}$); Thm 3.8 is proved in Thm E.34. □

4 Experiment

We evaluate MixOPD under two representative OPD formulations: standard OPD and EOPD. Student policies are initialized from Qwen3-VL-4B/8B without *cold start*, while the teacher model is fixed as Qwen3-VL-32B. Our training data is curated from *TreeVGR-RL-37K* Wang et al. (2025a) and *ZwZ-RL-VQA* Wei et al. (2026). Following recent work Yuan et al. (2026); Cai et al. (2026), we condition the teacher on highlighted evidence regions as visual privileges and ground-truth answers as logical privileges. Training is conducted on 16× H800 80GB GPUs. We compare MixOPD with recent alternatives and representative post-training baselines across both fine-grained and general-purpose visual tasks. Further implementation and evaluation details are provided in Appendices B and C.

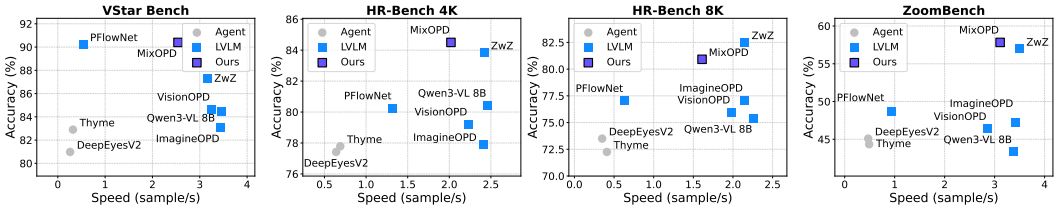

Figure 4: Performance–latency analysis of alternatives at the 8B scale across four benchmarks.

4.1 Comparison with SOTA LVLMs

Main Results. As shown in Table 1, MixOPD brings consistent gains over the base models across various visual tasks, which improves Qwen3-VL-4B/8B by 5.61*pp* and 4.99*pp*, respectively. MixOPD delivers an impressive performance–parameter trade-off. At the 4B scale, MixOPD reaches performance comparable to the 106B GLM-4.6V on $7/8$ benchmarks. At the 8B scale, MixOPD further closes the gap to private LVLMs and achieves results comparable to the 1T-parameter Kimi-K2.6 on most fine-grained benchmarks. MixOPD also compares favorably with strong vision-centric reasoning frameworks (*e.g.*, DeepEyes, Thyme), and distillation-based alternatives (*e.g.*, VisionOPD, ImagineOPD). Against these baselines under comparable model scales, MixOPD achieves stronger overall performance, especially on fine-grained visual tasks, empirically validating its competitiveness.

Performance–Latency Analysis. As shown in Figure 4, MixOPD achieves a favorable performance–latency trade-off across challenging benchmarks. Compared with agentic frameworks, MixOPD jointly localizes visual evidence and performs evidence-aware reasoning within a single generation, without additional inference-time costs, *e.g.*, search procedures or tool calls. In evaluation, this yields substantially lower latency yet even higher accuracy, corresponding to approximately $3\times$–$9\times$ faster inference. Each grounding span adds only a small number of output tokens while providing explicit visual anchors for fine-grained reasoning. Consequently, MixOPD also outperforms distillation-based baselines that internalize evidence selection implicitly at comparable inference speed.

4.2 Comparison with Post-Training Baselines

Table 2: Comparison with representative post-training baselines. 'method‡' uses *accuracy* and *dual IoU* as *verifiable rewards*, while 'method$^{\#}$' uses *Qwen3-VL 32B* as a *privileged teacher* (Appx. B.2). Spatial calibration improves its corresponding OPD or EOPD baseline across every reported metric.

Method	Fine-Grained Visual Tasks						General-Purpose Visual Tasks			
	mIoU	TreeB	VStar	Zoom	HR_{4K}	HR_{8K}	MME^{fw}_{CN}	MME^{fw}_{EN}	MMStar	CV-B
Qwen3-VL 4B	0.31	43.29	80.11	40.51	77.79	72.63	63.61	63.94	70.44	85.18
GRPO‡ Guo et al. (2025)	0.43	46.83	81.85	44.46	78.96	73.88	63.20	60.06	64.02	83.52
DAPO‡ Yu et al. (2026a)	0.44	47.00	79.05	45.13	79.88	75.25	65.50	66.30	66.04	85.73
OPD$^{\#}$ Lu & Lab (2025)	0.26	47.00	84.29	49.19	81.08	75.67	64.04	67.46	70.22	85.60
+ **Spatial Div** (Ours)	**0.45** +0.19	51.03 +4.03	89.70 +5.41	55.38 +6.19	81.42 +0.34	**79.38** +3.71	66.18 +2.14	68.67 +1.21	70.42 +0.20	87.91 +2.31
EOPD$^{\#}$ Jin et al. (2026)	0.39	49.79	89.18	53.49	81.88	78.25	64.96	67.69	70.29	88.43
+ **Spatial Div** (Ours)	**0.45** +0.06	**52.18** +2.39	**90.92** +1.74	**56.49** +3.00	**82.83** +0.95	78.79 +0.54	**67.65** +2.69	**68.84** +1.15	**71.58** +1.29	**89.37** +0.94
Qwen3-VL 8B	0.35	45.10	84.47	43.32	80.42	75.42	63.52	64.83	71.11	85.17
GRPO‡ Guo et al. (2025)	0.45	47.82	86.21	46.75	78.42	74.50	59.13	60.57	65.05	86.47
DAPO‡ Yu et al. (2026a)	**0.48**	48.73	88.31	44.34	78.38	76.00	66.09	63.44	66.42	85.90
OPD$^{\#}$ Lu & Lab (2025)	0.28	49.96	87.26	52.66	82.71	79.96	65.09	67.31	71.98	86.15
+ **Spatial Div** (Ours)	0.46 +0.18	53.25 +3.29	90.40 +3.14	57.83 +5.17	84.50 +1.79	80.92 +0.96	66.53 +1.44	68.27 +0.96	72.35 +0.37	87.39 +1.24
EOPD$^{\#}$ Jin et al. (2026)	0.42	51.52	89.70	54.36	83.25	80.96	66.46	67.87	72.24	86.20
+ **Spatial Div** (Ours)	**0.48** +0.06	**53.99** +2.47	**92.50** +2.80	**59.76** +5.40	**85.25** +2.00	**81.13** +0.17	**68.10** +1.64	**69.98** +2.11	**73.71** +1.47	**88.36** +2.16

As shown in Table 2, the spatially calibrated variants achieve the strongest overall performance across the evaluated benchmarks, involving both fine-grained visual understanding and general-purpose VQA. GRPO and DAPO substantially improve *mIoU* through accuracy-plus-dual-IoU rewards, but their reasoning gains are smaller and less consistent, suggesting that sparse sequence-level rewards do not couple localization and reasoning as effectively as dense teacher supervision. Standard OPD yields broader accuracy gains yet reduces *mIoU* at both scales, a pattern consistent with the pre-clipping serialization-sensitive coordinate pressure diagnosed in Lemma 3.6. EOPD improves both grounding and reasoning, consistent with its entropy-gated forward-KL term protecting retained teacher-supported modes that are under-covered by the student. Spatial calibration further improves every reported metric under both standard OPD and EOPD, with particularly pronounced gains on fine-grained tasks, where localized evidence plays a more central role. This empirical division of labor mirrors the theory: Theorem 3.8 establishes proposal-local scaling of the implemented local (sub)gradient according to geometry, while Corollary E.38 formalizes its composition with EOPD.

4.3 Ablation Study

Table 3: Ablation of spatial-search hyperparameters: the number V of returned teacher logits (*left*), position-wise search width M_c, and the number K of retained canonical boxes (*right*).

V	TreeB	VStar	Zoom	$\text{MME}^{\text{rw}}_{\text{EN}}$
MixOPD with standard OPD				
32	52.26	89.01	56.92	67.94
64	53.25	90.40	**57.83**	**68.27**
128	**53.33**	**91.10**	57.51	68.05
MixOPD with EOPD				
32	52.67	90.75	58.66	69.31
64	**53.99**	**92.50**	**59.76**	**69.98**
128	53.66	91.80	59.41	69.00

M_c	TreeB	VStar	Zoom	$\text{MME}^{\text{rw}}_{\text{EN}}$
MixOPD with standard OPD				
32	52.67	87.61	57.08	67.61
64	**53.25**	**90.40**	**57.83**	**68.27**
128	53.17	89.88	56.21	67.39
MixOPD with EOPD				
32	53.17	90.92	57.59	68.13
64	**53.99**	**92.50**	59.76	**69.98**
128	52.76	92.15	**60.28**	69.26

K	TreeB	VStar	Zoom	$\text{MME}^{\text{rw}}_{\text{EN}}$
MixOPD with standard OPD				
32	52.10	90.23	56.92	67.04
64	52.76	**90.58**	56.53	67.99
128	**53.25**	90.40	**57.83**	**68.27**
MixOPD with EOPD				
32	52.51	90.92	59.21	68.45
64	**54.16**	92.32	**60.00**	67.77
128	53.99	**92.50**	59.76	**69.98**

Spatial-search Hyperparameters. Table 3 examines how the three key hyperparameters of teacher-forced constrained search (§B.2) affect MixOPD. Aggressive truncation at $V = 32$ is consistently inferior to $V = 64$, while $V = 64$ outperforms $V = 128$ in 6/8 metric–formulation pairs at lower cost. At fixed $V = 64$ and $K = 128$, $M_c = 64$ performs best among the tested widths in 7/8 pairs. It retains useful partial coordinate-value paths after each digit extension; further widening mainly retains lower-probability, near-tied paths and yields limited gains. In particular, at fixed $V = M_c = 64$, the better of $K = 64$ and $K = 128$ outperforms $K = 32$ on every metric. This saturation follows from the finite-candidate analysis (Lemma E.31): V, M_c, and K fix which candidates are retained, while approximation quality tracks retained teacher mass, candidate validity, and the frozen-context factorization gap. Enlarging a cutoff displaces candidates and moves these three quantities in different directions, and $V = M_c = 64$ with $K = 128$ balances them.

Table 4: Ablation of teacher-side privilege under standard OPD (*left*) and EOPD (*right*), where `Triv` gives the teacher the same inputs as the student; $\texttt{Priv}_{\text{V}}$ highlights the refined evidence boxes in the teacher image, $\texttt{Priv}_{\text{T}}$ conditions the teacher on the ground-truth answer, and $\texttt{Priv}_{\text{all}}$ combines both privileges. Student inputs remain unprivileged in all settings.

	mIoU	TreeB	VStar	Zoom	$\text{MME}^{\text{rw}}_{\text{EN}}$
Teacher	0.37	46.83	88.13	49.94	66.97
Triv	$0.34_{-0.03}$	$46.75_{-0.08}$	$87.61_{-0.52}$	$46.87_{-3.07}$	$65.71_{-1.26}$
Priv$_{\text{V}}$	$0.42_{+0.05}$	$50.70_{+3.87}$	$88.83_{+0.70}$	$53.65_{+3.71}$	$68.15_{+1.18}$
Priv$_{\text{T}}$	$0.41_{+0.04}$	$50.04_{+3.21}$	$89.18_{+1.05}$	$54.63_{+4.69}$	$67.69_{+0.72}$
Priv$_{\text{all}}$	$\mathbf{0.45}_{+0.08}$	$\mathbf{53.25}_{+6.42}$	$\mathbf{90.40}_{+2.27}$	$\mathbf{57.83}_{+7.89}$	$\mathbf{68.27}_{+1.30}$

	mIoU	TreeB	VStar	Zoom	$\text{MME}^{\text{rw}}_{\text{EN}}$
Teacher	0.37	46.83	88.13	49.94	66.97
Triv	$0.37_{+0.00}$	$47.74_{+0.91}$	$87.96_{-0.17}$	$49.19_{-0.75}$	$67.34_{+0.37}$
Priv$_{\text{V}}$	$0.46_{+0.09}$	$51.19_{+4.36}$	$90.92_{+2.79}$	$58.46_{+8.52}$	$67.98_{+1.01}$
Priv$_{\text{T}}$	$0.42_{+0.05}$	$52.02_{+5.19}$	$92.15_{+4.02}$	$58.58_{+8.64}$	$68.34_{+1.37}$
Priv$_{\text{all}}$	$\mathbf{0.48}_{+0.11}$	$\mathbf{53.99}_{+7.16}$	$\mathbf{92.50}_{+4.37}$	$\mathbf{59.76}_{+9.82}$	$\mathbf{69.98}_{+3.01}$

Teacher-side Privilege. Table 4 shows that visual and logical privileges provide complementary benefits. Visual privilege gives the teacher direct access to task-relevant regions when scoring student rollouts and constructing the searched box distribution. Accordingly, $\texttt{Priv}_{\text{V}}$ yields the larger *mIoU* gain under both formulations, supporting the value of evidence-focused teacher guidance for spatial calibration. $\texttt{Priv}_{\text{T}}$ conditions the teacher on the ground-truth answer, refining its token distribution toward decision-relevant reasoning; its larger accuracy benefit under EOPD suggests that EOPD's entropy-gated forward-KL term benefits from this refined distribution. Combining both privileges consistently performs best. Because only the teacher receives privileged inputs during training, these results show that complementary visual and logical information can be transferred into a student that operates on standard inputs and even outperforms the teacher evaluated under those inputs.

4.4 In-depth Analysis

Figure 5: Token-level entropy distributions (*left*); the fraction and composition of FKL-activated tokens across threshold η_H (*right*).

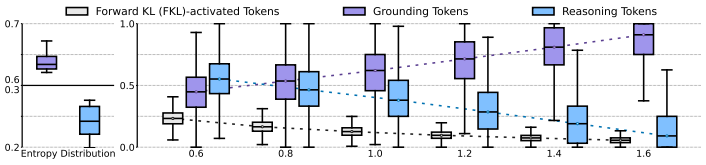

Table 5: Results of EOPD under various threshold η_H.

η_H	*mIoU*	TreeB	VStar	Zoom
0.6	0.37	49.55	87.78	52.62
0.8	0.38	50.04	88.48	54.04
1.0	0.40	50.04	88.66	**54.36**
1.2	0.42	**51.52**	**89.70**	**54.36**
1.4	**0.43**	**51.52**	89.18	54.12
1.6	0.42	51.44	88.31	52.94

Effectiveness of EOPD in Visual Reasoning. EOPD shows clear and systematic gains over standard OPD in the preceding experiments; we therefore analyze why it is particularly effective for vision-centric reasoning. EOPD benefits this setting because its entropy-gated forward-KL term preferentially supplies mode-covering supervision at high-uncertainty grounding decisions that select the visual evidence used by subsequent reasoning. Figure 5 demonstrates that grounding tokens have

substantially higher teacher entropy than reasoning tokens As η_H rises from 0.6 to 1.6, the overall FKL-activated fraction falls from $\sim 22\%$ to $\sim 5\%$, while the grounding share among activated tokens rises from $\sim 45\%$ to $\sim 90\%$, progressively concentrating the remaining supervision on grounding decisions. Accordingly, the best operating threshold in our visual-reasoning setting shifts from the language-centric default $\eta_H = 0.8$ to 1.2: Table 5 shows simultaneous gains in mIoU and accuracy, suggesting that grounding-focused supervision improves both evidence localization and subsequent reasoning. At $\eta_H = 1.6$, the activated set narrows further and VStar and Zoom decline, making 1.2 the best overall balance between grounding selectivity and token coverage. The forward-KL analysis explains this pattern through its non-vanishing recovery gradients and coverage pressure on retained teacher-supported alternatives under-covered by the student (Lem. E.26; Cor. E.25). Furthermore, EOPD and MixOPD compose effectively because they improve complementary aspects of OPD supervision. Adding Spatial Divergence to EOPD improves every reported metric at both model scales, while the controlled ablation in Appx. D.1 shows that entropy gating mainly strengthens reasoning-oriented metrics and proposal-level spatial calibration yields larger grounding and fine-grained gains. EOPD broadens teacher-supported coverage at uncertain token positions, whereas MixOPD uses complete-proposal geometry to calibrate the strength of the complete local surrogate on grounding spans. Cor. E.38 formalizes their composition in the implemented objective.

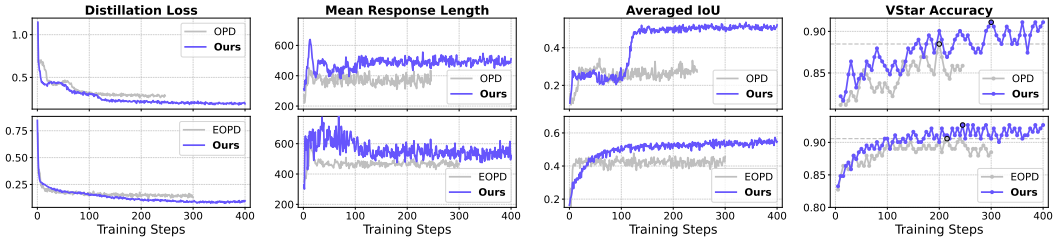

Figure 6: Training dynamics of MixOPD on Qwen3-VL-8B under standard OPD and EOPD formulations, where *Averaged IoU* measures the spatial consistency between student-predicted grounding boxes and ground-truth annotations across all rollouts, using sequence-level dual IoU.

Training Dynamics under Different Formulations. Figure 6 reveals a separation between token-level optimization and spatial learning. Under standard OPD, the sampled token-level surrogate continues to decrease while averaged IoU remains low, showing that better optimization of the sampled objective need not improve localization. Spatial calibration changes this trajectory: after a brief warm-up, IoU rises sharply to a higher plateau alongside continued gains in VStar accuracy. A plausible interpretation is an on-policy bootstrap: as parseable grounding proposals become more frequent, complete-proposal geometry can reweight the local surrogate at more coordinate-bearing positions. This is consistent with Theorem E.34, which attenuates local gradients on valid proposals while retaining a separation-dependent fraction for spatially separated errors. EOPD improves IoU and accuracy earlier, consistent with its entropy-gated forward-KL term supplying recovery signals to retained teacher-supported modes under-covered by the student, but its uncalibrated IoU soon saturates. Spatial calibration continues to improve both metrics, indicating that token-level uncertainty alone does not determine the geometric validity of a complete proposal. Both calibrated variants settle into longer yet stable responses after an early transient, alongside sustained improvements in performance. Overall, EOPD improves mode coverage in token space, whereas MixOPD calibrates coordinate-level learning through complete-proposal geometry in image space.

5 Conclusion

We study OPD for vision-centric reasoning and identify a fundamental mismatch between token-level distribution matching and spatial grounding: exact coordinate serializations remain distinct outcomes even when they localize the same visual evidence, which can induce serialization-sensitive pressure on spatially valid proposals. To address this, we present MixOPD, which uses the *Spatial Divergence* between each on-policy student proposal and a teacher-supported evidence distribution to calibrate the implemented local distillation surrogate at coordinate-bearing positions, provably attenuating local gradient magnitudes for valid proposals while retaining a separation-dependent fraction for spatially separated errors. Across Qwen3-VL-4B/8B, MixOPD consistently improves grounding and reasoning when instantiated with either standard OPD or EOPD on various visual tasks, while exhibiting stable training dynamics without adding inference-time overheads.

Reproducibility Statement

MixOPD is reproducible entirely from public resources. The student and teacher use publicly available Qwen3-VL-Instruct checkpoints, and the training corpus is reconstructed exclusively from TreeVGR-RL-37K and ZwZ-RL-VQA, without private, proprietary, or newly collected data. Appendix B documents the complete training pipeline: §B.1 details leakage auditing, dual-hash deduplication, construction of the ∼145K-example corpus, open-source DeepScan Li et al. (2026b) evidence refinement, and generation of visual and logical teacher privileges, with accompanying code that reconstructs the corpus from its original sources; §B.2 specifies teacher-forced constrained coordinate search, including position-wise scoring and filtering; §B.3 defines the exact objectives and training recipes, including global valid-token normalization; §B.4 reports the hyperparameters, spatial-search configuration, and VeRL-aligned GRPO and DAPO settings; §B.5 provides the common grounded-reasoning prompt; and §B.6 gives complete pseudocode for MixOPD instantiated with standard OPD and additive EOPD, covering span parsing, distribution construction, spatial calibration, gradient boundaries, fallback behavior, and computational complexity. Appendix C specifies the benchmarks and metrics (§C.1), baselines (§C.2), and unified evaluation protocol (§C.3), explicitly distinguishing official-checkpoint evaluation from backbone-matched reproduction, while Appendix D.2 defines the paired significance tests and cross-seed replication analysis. Appendix E provides the complete definitions, assumptions, proofs, finite-candidate analysis, and implementation-faithful local guarantees. The supplement contains all experimental code and configurations; we will additionally release the processed annotations, trained checkpoints, and a step-by-step reproduction guide. Together, these materials enable independent reproduction of all MixOPD and reproduced-baseline results without private data or unpublished author-side components.

References

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

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report, 2025. *URL https://arxiv.org/abs/2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Yishuo Cai, Jiahui Liu, Yuanxin Liu, Haobo Deng, Linli Yao, Yuhao Zheng, Kun Ouyang, Zhimo Li, Ziyue Wang, Xu Sun, et al. Thinking without images: Internalizing visual manipulation with on-policy self-distillation. *arXiv preprint arXiv:2606.08719*, 2026.

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

Yong Xien Chng, Tao Hu, Wenwen Tong, Xueheng Li, Jiandong Chen, Haojia Yu, Jiefan Lu, Hewei Guo, Hanming Deng, Chengjun Xie, et al. Sensenova-mars: Empowering multimodal agentic reasoning and search via reinforcement learning. *arXiv preprint arXiv:2512.24330*, 2025.

DeepMind. Gemini 3.1 pro. https://deepmind.google/models/model-cards/gemini-3-1-pro/, 2026a.

DeepMind. Gemini 3.5 flash. https://deepmind.google/models/model-cards/gemini-3-5-flash/, 2026b.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*, 2025a.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025b.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

Kimi Team. Kimi k2.6: From code to creation, from one to many. https://www.kimi.com/ai-models/kimi-k2-6/, 2026. Accessed: 2026-06-22.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Aristotelis Lazaridis, Dylan Bates, Aman Sharma, Brian King, Vincent Lu, and Jack FitzGerald. Edge-opd: Internalizing privileged context with evidence guided on-policy distillation. *arXiv preprint arXiv:2605.23493*, 2026.

Yangfu Li, Yuning Gong, Hongjian Zhan, Teng Li, Yuanhuiyi Lyu, Tianyi Chen, Qi Liu, Ziyuan Huang, Zhihang Zhong, Dandan Zheng, et al. Perceptual flow network for visually grounded reasoning. *arXiv preprint arXiv:2605.02730*, 2026a.

Yangfu Li, Hongjian Zhan, Jiawei Chen, Yuning Gong, Qi Liu, and Yue Lu. Deepscan: A training-free framework for visually grounded reasoning in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19155–19164, 2026b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/, 2024.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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

Ruijie Lu, Yiyang Ma, Xiaokang Chen, Lingxiao Luo, Zhiyu Wu, Zizheng Pan, Xingchao Liu, Yutong Lin, Hao Li, Wen Liu, et al. Thinking with visual primitives. *DeepSeek-AI technical report*, 2026.

Run Luo, Lu Wang, Wanwei He, Longze Chen, Jiaming Li, and Xiaobo Xia. Gui-r1: A generalist r1-style vision-language action model for gui agents. *arXiv preprint arXiv:2504.10458*, 2025.

OpenAI. Introducing gpt-5.2. https://openai.com/index/introducing-gpt-5-2/, 2025.

OpenAI. Introducing gpt-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

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

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. *arXiv preprint arXiv:2507.07999*, 2025a.

Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, et al. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*, 2025b.

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

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

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13084–13094, 2024.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026a.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11704–11715, 2026b.

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

Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, et al. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 5, 2025.

Junkai Zhang, Yihe Deng, Kai-Wei Chang, and Wei Wang. Thinking with visual grounding. *arXiv preprint arXiv:2606.16122*, 2026a.

Ke Zhang, Yunjie Tian, DongDi Zhao, Yijiang Li, Yuanye Liu, Vishal M Patel, and Di Fu. On-policy distillation with best-of-n teacher rollout selection. *arXiv preprint arXiv:2605.09725*, 2026b.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025a.

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

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

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

APPENDIX OVERVIEW

The appendices are organized to support three complementary reading paths. Readers interested in the methodological rationale can begin with Appendix A and then consult Appendix E for the complete formal analysis. Readers seeking to reproduce the study can proceed directly to Appendices B and C, which document the complete training, implementation, and evaluation configurations. Readers interested in additional empirical evidence can jump to Appendix D. The remaining statements are collected in Appendix F.

Methodological analyses and validation. Appendix A examines the three methodological questions most directly associated with spatial calibration. §A.1 relates the analytic calibration field to 160K successfully searched grounding spans observed during training. It shows that raw coordinate pressure does not reliably order spatial agreement, whereas MixOPD introduces geometry-conditioned retention and strongly contracts the residual pressure on highly aligned proposals. §A.2 compares the mode, uniform, and probability-weighted expectation formulations under annotation-matched diagnostics. Mode and uniform exhibit complementary endpoint failures, while expectation provides conditional-rank robustness without requiring an annotation-dependent oracle to switch between them; the analysis further shows that teacher probabilities carry a weak but systematic signal of proposal geometry. Finally, §A.3 evaluates teacher-forced constrained coordinate search against candidate-specific autoregressive rescoring and constrained beam search. The results show that TFCS preserves the scalar spatial coefficient substantially more closely than the underlying candidate-law differences would suggest, while eliminating candidate-specific teacher decoding and reducing the measured search overhead by more than three orders of magnitude.

Training and implementation details. Appendix B provides the complete training-side specification. §B.1 describes the leakage audit, construction of the approximately 145K-example training corpus, DeepScan-based evidence refinement, and the visual and logical teacher privileges. §B.2 gives the exact TFCS candidate-construction procedure, including position-wise truncation, tie-aware constrained merges, geometric admissibility, deterministic tie-breaking, and fallback behavior. §B.3 defines the implemented objectives and training recipes for GRPO, DAPO, standard OPD, EOPD, and MixOPD, while §B.4 reports all shared and method-specific hyperparameters. §B.5 provides the common grounded reasoning prompt. §B.6 concludes with the complete MixOPD pseudocode, gradient boundaries, and computational analysis, showing that spatial calibration requires no additional model evaluation and does not change the leading neural-compute order of OPD.

Evaluation setup. Appendix C specifies how all reported results are obtained. §C.1 introduces the general-purpose and fine-grained benchmarks together with their accuracy and localization metrics. §C.2 describes the agentic, proprietary, open-source, distillation-based, and post-training baselines used for comparison. §C.3 then details the unified inference pipeline, official-checkpoint evaluation, backbone-matched VisionOPD reproduction, decoding configuration, two-stage answer scoring, and three-seed evaluation protocol.

Additional empirical results. Appendix D supplements the main experiments from three perspectives. §D.1 examines teacher scale, invalid-span handling, and the separate choices of where unreliable supervision is localized and how it is recalibrated. These results support span-level degradation over sample-level removal and further distinguish the complementary roles of entropy-aware forward KL and spatial-divergence calibration. §D.2 reports the raw paired correctness counts, exact two-sided McNemar tests, panel-wise Holm correction, and descriptive cross-seed replication profiles for both competitive alternatives and the corresponding OPD/EOPD baselines. §D.3 presents representative rollouts and per-position teacher-entropy traces, illustrating explicit and inspectable grounding, question-conditioned visual search, localized uncertainty within grounding spans, and the complementary scopes of token-level EOPD and span-level MixOPD.

Theoretical foundations and guarantees. Appendix E gives the complete theoretical development. §E.1 constructs normalized population coordinate-decision laws and defines the implemented OPD–EOPD surrogate. §E.2 decomposes coordinate-level KL into evidence-level disagreement and a within-class residual over exact coordinate realizations. §E.3 states the separate assumptions used for dilution, teacher reliability, and EOPD mode coverage. §E.4 proves the $\Theta(\log M)$ exact-coordinate pressure induced by valid-realization dilution and characterizes its aggregate entropy signature. §E.5 establishes the coverage barrier and missing-mode recovery role of the additive forward-KL component. §E.6 develops the spatial coefficient, finite-candidate approximation, frozen-context

factorization gap, and robust selectivity bounds for the practical teacher law. Finally, §E.7 proves the implementation-faithful geometry-selective guarantee: the same detached coefficient scales the complete local OPD/EOPD surrogate, suppressing valid proposals according to ε_{eff} while retaining a Δ_{eff}-dependent fraction on spatially separated errors. These are local supervision guarantees for fixed rollouts, rather than claims of global convergence or monotone geometric improvement.

Statements. For completeness and transparency, Appendix F summarizes the §F.1 limitations and future work, discusses the §F.2 broader impacts, and provides the §F.3 LLM-use declaration.

A Methodological Analyses and Validation

A.1 Geometry-Selective Calibration of Realized Coordinate Pressure

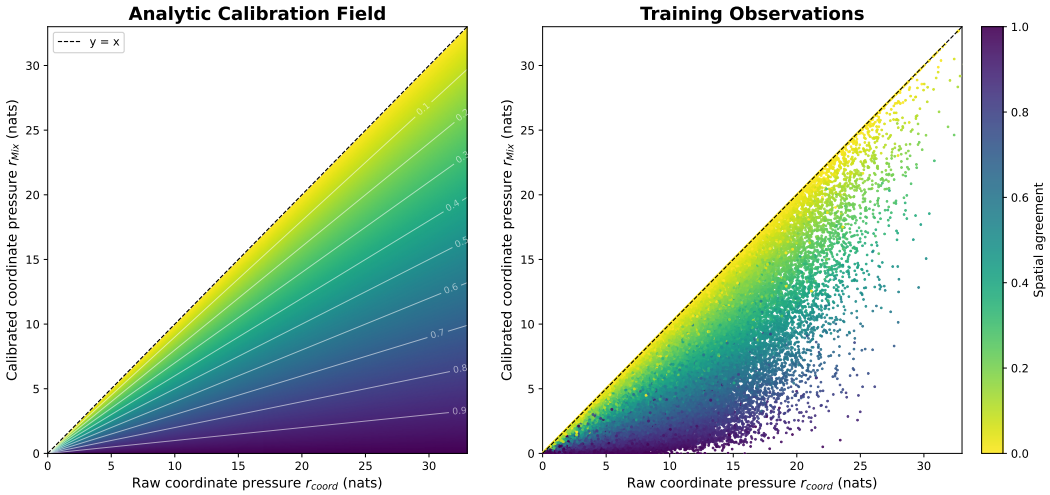

Figure 7: Geometry-selective calibration of realized coordinate pressure in MixOPD. *Left:* The analytic field $r_j^{\text{Mix}} = \widehat{\omega}_{j,K} r_j^{\text{coord}}$. Each ray has slope $\widehat{\omega}_{j,K}$, while its color and contour denote the complementary expected spatial agreement $1 - \widehat{\omega}_{j,K} = \mathbb{E}_{b \sim \widehat{P}^{\mathcal{B}}_{\text{tea},K}}[\text{IoU}(b_j, b)]$; $y = x$ denotes no attenuation. *Right:* 160K realized grounding spans sampled throughout training. The horizontal axis is the unclipped coordinate log-density ratio r_j^{coord}, the vertical axis is r_j^{Mix} computed with the actual coefficient used by MixOPD, and its color denotes $\text{IoU}(b_j, \widehat{b}_j^{\text{MAP}})$. The observation demonstrates that *raw coordinate pressure does not reliably order spatial agreement, whereas calibration introduces geometry-conditioned retention and strongly contracts the high-alignment regime.*

Probing setup. To quantify how strongly the geometry-selective mechanism in Theorem 3.8 acts on the successfully supervised portion of the realized training distribution, we instrument MixOPD and periodically record the coordinate pressures of 160K realized, *well-formed* grounding spans for which teacher-forced constrained search *succeeds* and yields a *nonempty* practical teacher box law. Fallback spans are excluded from this diagnostic cohort. For each span b_j, the raw diagnostic is the unclipped coordinate log-density ratio from Lemma 3.6,

$$r_j^{\text{coord}} = \sum\nolimits_{t \in \mathcal{I}(b_j)} \log \frac{\pi_{\theta_{\text{old}}}(x_t \mid \mathbf{c}_t)}{\pi_{\text{tea}}(x_t \mid \widetilde{\mathbf{c}}_t)} = -\sum\nolimits_{t \in \mathcal{I}(b_j)} A_t^{\text{raw}}.$$

We additionally record its auxiliary calibrated counterpart $r_j^{\text{Mix}} = \widehat{\omega}_{j,K} r_j^{\text{coord}}$, using the same detached coefficient $\widehat{\omega}_{j,K}$ applied by MixOPD during training. Here

$$\widehat{\omega}_{j,K} = 1 - \mathbb{E}_{b \sim \widehat{P}^{\mathcal{B}}_{\text{tea},K}(\cdot \mid \widetilde{\mathbf{c}}_{a_j})} \left[\text{IoU}(b_j, b)\right].$$

For an interpretable single-box diagnostic, we identify the highest-probability teacher proposal

$$\widehat{b}_j^{\text{MAP}} \in \operatorname*{arg\,max}_{b \in \mathcal{B}^K_{\text{tea}}(\widetilde{\mathbf{c}}_{a_j})} \widehat{P}^{\mathcal{B}}_{\text{tea},K}\left(b \mid \widetilde{\mathbf{c}}_{a_j}\right)$$

and record $\mathrm{IoU}(b_j, \widehat{b}_j^{\mathrm{MAP}})$. Following Remark E.16, r_j^{coord} is the unclipped sampled log-density ratio. It diagnoses the coordinate-level source pressure that enters OPD; under the bounded-dilution conditions of Lemma 3.6, this pressure grows as $\Theta(\log M)$ along problem families with increasing numbers of valid coordinate realizations.

Claim 1: MixOPD restores geometry-consistent credit assignment that standard OPD lacks. Figure 7 directly exposes the proposal-level credit-assignment mismatch of standard OPD and how spatial calibration corrects it. In the empirical panel, vertical slices contain grounding spans with similar raw coordinate pressure r_j^{coord} but widely different spatial agreement: proposals that are well aligned with the teacher evidence distribution can receive nearly the same raw pressure as substantially misaligned ones. Hence, the token-level OPD signal does not reliably order grounding behavior by geometric quality. The MAP-box IoU used for visualization is a faithful proxy for the distribution-level agreement entering MixOPD, with $\mathrm{corr}\Big(\mathrm{IoU}(b_j, \widehat{b}_j^{\mathrm{MAP}}),\, 1 - \widehat{\omega}_{j,K}\Big) = 0.976$ over the recorded spans. Spatial calibration introduces the missing ordering explicitly. Let

$$u_j := 1 - \widehat{\omega}_{j,K} = \mathbb{E}_{\widetilde{b} \sim \widehat{P}^{\mathcal{B}}_{\mathrm{tea},K}(\cdot \mid \widetilde{\mathbf{c}}_{a_j})} \Big[\mathrm{IoU}(b_j, \widetilde{b}) \Big]$$

denote the expected spatial agreement. For any fixed raw pressure $r > 0$,

$$r^{\mathrm{Mix}}(r, u) = (1 - u) r, \qquad \frac{\partial r^{\mathrm{Mix}}(r, u)}{\partial u} = -r < 0.$$

Thus, among proposals receiving the same raw OPD pressure, better spatial agreement necessarily receives less retained adverse pressure under MixOPD. The analytic field in Figure 7 visualizes exactly this monotone rotation: increasing agreement continuously moves the calibrated pressure away from the no-attenuation diagonal and toward the horizontal axis. This pressure-level relation is the scalar counterpart of Theorem 3.8, which shows that the same detached coefficient rescales the implemented local surrogate and semigradient without reversing their base direction. MixOPD therefore supplies the image-space ordering absent from standard OPD, explaining why the latter can continue improving its sampled token-level objective while localization remains poor, whereas spatial calibration assigns progressively weaker corrective pressure to geometrically better proposals.

Claim 2: MixOPD contracts realization-sensitive pressure on spatially aligned proposals while retaining corrective pressure on misaligned ones. To be specific, Figure 7 further reveals a distinct stability advantage. Highly aligned proposals remain broadly dispersed along the raw-pressure axis, showing that geometrically similar behavior can incur substantially different coordinate-level OPD pressure solely because of its exact token realization. This is consistent with the valid-realization dilution mechanism of Lemma 3.6, under which exact-coordinate pressure can grow as $\Theta(\log M)$ as probability mass is distributed across multiple spatially equivalent realizations. After calibration, the same high-alignment observations collapse into a much narrower region near the horizontal axis. This contraction is deterministic once the spatial coefficient is fixed:

$$\mathrm{Std}\big(r_j^{\mathrm{Mix}} \mid \widehat{\omega}_{j,K} = \omega\big) = \omega \ \mathrm{Std}\big(r_j^{\mathrm{coord}} \mid \widehat{\omega}_{j,K} = \omega\big)\,,$$

and

$$\mathrm{Var}\big(r_j^{\mathrm{Mix}} \mid \widehat{\omega}_{j,K} = \omega\big) = \omega^2 \ \mathrm{Var}\big(r_j^{\mathrm{coord}} \mid \widehat{\omega}_{j,K} = \omega\big)\,.$$

In particular,

$$1 - \widehat{\omega}_{j,K} \geq 0.9 \quad \Longrightarrow \quad \big|r_j^{\mathrm{Mix}}\big| \leq 0.1 \, \big|r_j^{\mathrm{coord}}\big|\,, \qquad \big(r_j^{\mathrm{Mix}}\big)^2 \leq 0.01 \, \big(r_j^{\mathrm{coord}}\big)^2\,.$$

The 0.9-agreement contour therefore forms the envelope $r_j^{\mathrm{Mix}} = 0.1\, r_j^{\mathrm{coord}}$, quantitatively explaining why highly aligned proposals can exhibit large realization-dependent variation before calibration yet remain tightly concentrated afterward. Crucially, this is not uniform attenuation: spatially discrepant proposals have larger $\widehat{\omega}_{j,K}$ and therefore retain a larger fraction of their corrective signal, consistent with the valid-proposal suppression and separated-error retention guarantees of Theorem 3.8. MixOPD consequently acts as geometry-adaptive local annealing, suppressing residual exact-token pressure and its variability once grounding is already aligned while preserving stronger correction where genuine spatial discrepancy remains. This selective contraction provides a direct mechanism for the more stable grounding optimization observed during training.

A.2 Robust Spatial Calibration under Uncertain Teacher Supervision

Formulations. For each successfully searched, well-formed grounding span $\tau(b_j)$, MixOPD constructs a scalar coefficient $\widehat{\omega}_{j,K}$, applied at the coordinate-bearing positions $t \in \mathcal{I}(b_j)$. Ideally, this coefficient quantifies the geometric inconsistency between b_j and the visual evidence supporting that reasoning step. However, such span-aligned evidence is generally not directly observable: reference annotations may be incomplete, and arbitrary on-policy spans cannot in general be matched one-to-one to the ground-truth boxes. MixOPD therefore uses the top-K teacher proposals $\mathcal{B}^K_{\mathrm{tea}}(\widetilde{\mathbf{c}}_{a_j})$ as a practicable surrogate for this evidence. This gives three formulations of the *student-to-evidence risk*:

$$\text{Mode:}\quad \widehat{\omega}^{\mathrm{mode}}_{j,K} = 1 - \mathrm{IoU}\Big(b_j, \widehat{b}^{\mathrm{MAP}}_j\Big)\,,\ \text{where } \widehat{b}^{\mathrm{MAP}}_j \in \arg\max_{b \in \mathcal{B}^K_{\mathrm{tea}}(\widetilde{\mathbf{c}}_{a_j})} \widehat{P}^{\mathcal{B}}_{\mathrm{tea},K}\big(b \mid \widetilde{\mathbf{c}}_{a_j}\big)\,,$$

$$\text{Uniform:}\quad \widehat{\omega}^{\mathrm{unif}}_{j,K} = \mathbb{E}_{b\sim U^{\mathcal{B}}_{j,K}}[1 - \mathrm{IoU}(b_j, b)]\,,\ \text{where } U^{\mathcal{B}}_{j,K} \text{ is the uniform law over } \mathcal{B}^K_{\mathrm{tea}}(\widetilde{\mathbf{c}}_{a_j}),$$

$$\text{Expectation:}\quad \widehat{\omega}^{\mathrm{exp}}_{j,K} = \mathbb{E}_{b\sim \widehat{P}^{\mathcal{B}}_{\mathrm{tea},K}(\cdot\mid\widetilde{\mathbf{c}}_{a_j})}[1 - \mathrm{IoU}(b_j, b)] \equiv \widehat{\omega}_{j,K}.$$

Mode collapses the retained teacher law to its MAP point mass, whereas uniform flattens it over the retained support. Expectation instead preserves both the support and its normalized teacher mass.

Probing setup. We seek to characterize when the *probability-weighted* expectation yields a more faithful estimate of the *student-to-evidence risk* than mode or uniform, and which properties of the retained teacher distribution accompany this advantage. We therefore distinguish mechanism-side selectivity from outcome-side formulation fidelity. The former captures geometric selectivity across the retained proposals, probabilistic selectivity of the teacher law, and the alignment between them; the latter uses matched annotations to quantify each formulation's calibration error and the expectation formulation's advantage over its alternatives. This separation enables us to probe how these selectivity factors relate to the regimes in which the expectation formulation is superior.

Geometric & Probabilistic Selectivity. Let $K_j := \big|\mathcal{B}^K_{\mathrm{tea}}(\widetilde{\mathbf{c}}_{a_j})\big|$ denote the actual retained support size, where $K_j \geq 2$ for every eligible span, and let $\{b_{j,k}\}_{k=1}^{K_j}$ be the retained teacher proposals. We define

$$q_{j,k} := \widehat{P}^{\mathcal{B}}_{\mathrm{tea},K}\big(b_{j,k} \mid \widetilde{\mathbf{c}}_{a_j}\big)\,, \qquad r_{j,k} := 1 - \max_{b^\star \in \mathcal{B}^\star_j} \mathrm{IoU}(b_{j,k}, b^\star),$$

where $\mathcal{B}^\star_j$ is the set of available reference annotations and is *used only for post-hoc diagnosis*. Let m_j denote the deterministic MAP index used in the analysis, so that $b_{j,m_j} = \widehat{b}^{\mathrm{MAP}}_j$; ties are resolved by the retained proposal order. We consider three properties of the retained teacher law:

$$S^{\mathrm{geo}}_j := \mathrm{Std}_{k\sim U_j}[r_{j,k}], \quad S^{\mathrm{prob}}_j := \frac{K_j}{\sqrt{K_j - 1}}\,\mathrm{Std}_{k\sim U_j}[q_{j,k}], \quad \rho^{\mathrm{align}}_j := -\,\mathrm{Corr}_{k\sim U_j}(q_{j,k}, r_{j,k}),$$

where U_j is uniform over the retained indices $[K_j]$. The alignment coefficient is undefined when either standard deviation is numerically zero; such degenerate spans are excluded from conditional alignment statistics. Here, S^{geo}_j quantifies *geometric selectivity*: the extent to which the reference annotations geometrically favor some retained teacher proposals over others, as reflected by variation in their annotation-relative errors. In parallel, S^{prob}_j quantifies *probabilistic selectivity*: the extent to which the teacher assigns greater probability mass to some retained proposals than to others. Finally, ρ^{align}_j bridges the geometry and probability: it measures whether proposals geometrically favored by the reference annotations receive greater teacher probability mass and, equivalently, whether teacher-preferred proposals are geometrically closer to the annotations. The aligned selection gain is defined for every eligible span by $A^{\mathrm{align}}_j := \mathbb{E}_{k\sim U_j}[r_{j,k}] - \mathbb{E}_{k\sim q_j}[r_{j,k}]$. For nondegenerate spans, it further admits the exact factorization

$$A^{\mathrm{align}}_j = \sqrt{K_j - 1}\, S^{\mathrm{geo}}_j S^{\mathrm{prob}}_j \rho^{\mathrm{align}}_j. \tag{4}$$

The geometric and probabilistic selectivity determine how consequential soft selection can be, whereas the alignment coefficient determines whether the resulting probability weighting is beneficial.

Soft- and hard-selection gains. To place expectation and mode on an annotation-relative scale, let $\bar{r}_j := \mathbb{E}_{k\sim U_j}[r_{j,k}]$ denote the uniform candidate risk and define the hard MAP-selection gain

$$A^{\mathrm{MAP}}_j := \bar{r}_j - r_{j,m_j}.$$

The corresponding annotation-side candidate risks induced by the three aggregation laws are

$$\underbrace{\mathbb{E}_{k\sim U_j}[r_{j,k}]}_{\text{uniform}} = \bar{r}_j, \qquad \underbrace{\mathbb{E}_{k\sim q_j}[r_{j,k}]}_{\text{expectation}} = \bar{r}_j - A_j^{\text{align}}, \qquad \underbrace{r_{j,m_j}}_{\text{mode}} = \bar{r}_j - A_j^{\text{MAP}}.$$

Consequently, on this annotation-side candidate-risk scale, uniform, expectation, and mode realize gains 0, A_j^{align}, and A_j^{MAP}, respectively, relative to the *uniform aggregation*. Unlike A_j^{align}, which measures the gain from preserving the full probability–geometry pairing, A_j^{MAP} measures the gain from hard MAP selection and depends on the relative geometric quality of the selected proposal.

Annotation-relative formulation diagnostics. Using the available annotations as a post-hoc reference, we define the annotation-relative errors of the realized student proposal and the teacher MAP proposal:

$$d_j^{\text{stu}} := 1 - \max_{b^\star \in \mathcal{B}_j^\star} \text{IoU}(b_j, b^\star), \qquad d_j^{\text{tea}} := 1 - \max_{b^\star \in \mathcal{B}_j^\star} \text{IoU}(\widehat{b}_j^{\text{MAP}}, b^\star) = r_{j,m_j},$$

and then compare expectation against the other two formulations:

$$e_j^g := \left| \widehat{\omega}_{j,K}^g - d_j^{\text{stu}} \right|, \quad g \in \{\text{exp}, \text{mode}, \text{unif}\}, \qquad \delta_j^f := e_j^f - e_j^{\text{exp}}, \quad f \in \{\text{mode}, \text{unif}\}.$$

Here, $\delta_j^f > 0$ indicates lower annotation-relative calibration error for expectation, whereas $\delta_j^f < 0$ favors the corresponding alternative. Importantly, d_j^{tea} is the realized annotation-relative MAP error; it measures neither teacher confidence nor probabilistic selectivity.

Annotation matching and analysis set. All diagnostics require reference annotations aligned with each realized grounding span. However, the approximately 160K spans recorded from arbitrary on-policy rollouts do not have a one-to-one correspondence with the available annotations, and some do not express grounding intent toward any annotated evidence. We therefore use a VLM-based judge conditioned on the input image and question, together with the reasoning context surrounding each span, to determine whether the span targets one or more annotated evidence regions and, if so, to identify the corresponding reference-annotation set $\mathcal{B}_j^\star$. After this semantic matching and diagnostic validity filtering, we retain 69,627 eligible spans from 29,199 image clusters for all subsequent analyses. This matching procedure is used only for post-hoc diagnosis and does not enter training.

Claim 1: Mode and uniform offer complementary advantages, whereas expectation provides a robust surrogate without oracle switching. Fig. 8 connects mechanism-side selectivity to outcome-side formulation fidelity. Panels **(a)–(c)** relate this mechanism to the empirical excess calibration errors while varying the reliability of the hard MAP choice. Panels **(d)–(f)** then express the same competition through the soft- and hard-selection gains relative to their common uniform baseline, focusing on the low-student-error regime in which annotation-side candidate risks most directly reflect coefficient fidelity. The three rows therefore distinguish whether selection can matter, which endpoint fails under the realized conditions, and why the formulation ranking changes across regimes.

$\diamondsuit$ *Mode and uniform occupy complementary endpoint regimes.* At high geometric selectivity, Panels **(a)** and **(b)** exhibit a clear reversal with MAP reliability. When d_j^{tea} is small, the upper-right cells have $\delta_j^{\text{mode}} < 0$ but $\delta_j^{\text{unif}} > 0$: mode benefits from the reliable MAP proposal, whereas uniform dilutes that informative selection. As d_j^{tea} increases, mode becomes increasingly unfavorable, while uniform is favored relative to expectation over much of the geometrically selective support. Panel **(c)** correspondingly switches from uniform being worst in the reliable, highly selective corner to mode being worst over the broader unreliable-MAP region. By contrast, formulation differences are small or statistically unresolved across much of the low-S_j^{geo} region, where the retained proposals have similar annotation-relative quality. These observations expose the complementary assumptions of the two endpoints: mode succeeds when hard MAP selection identifies the right proposal, whereas uniform avoids committing all mass to an unreliable MAP but discards useful selection when the MAP is reliable. Because this annotation-relative reliability regime is unavailable during training, neither endpoint provides a regime-independent surrogate for MixOPD.

$\diamondsuit$ *A soft–hard gain-space explanation of the endpoint complementarity.* In the idealized common-reference limit with $d_j^{\text{stu}} = 0$, the lines $A_j^{\text{align}} = 0$, $A_j^{\text{MAP}} = 0$, and $A_j^{\text{align}} = A_j^{\text{MAP}}$ define the pairwise equal-risk boundaries for expectation versus uniform, mode versus uniform, and expectation versus mode, respectively, providing a unified gain-space account of the endpoint reversals. Panel **(d)**

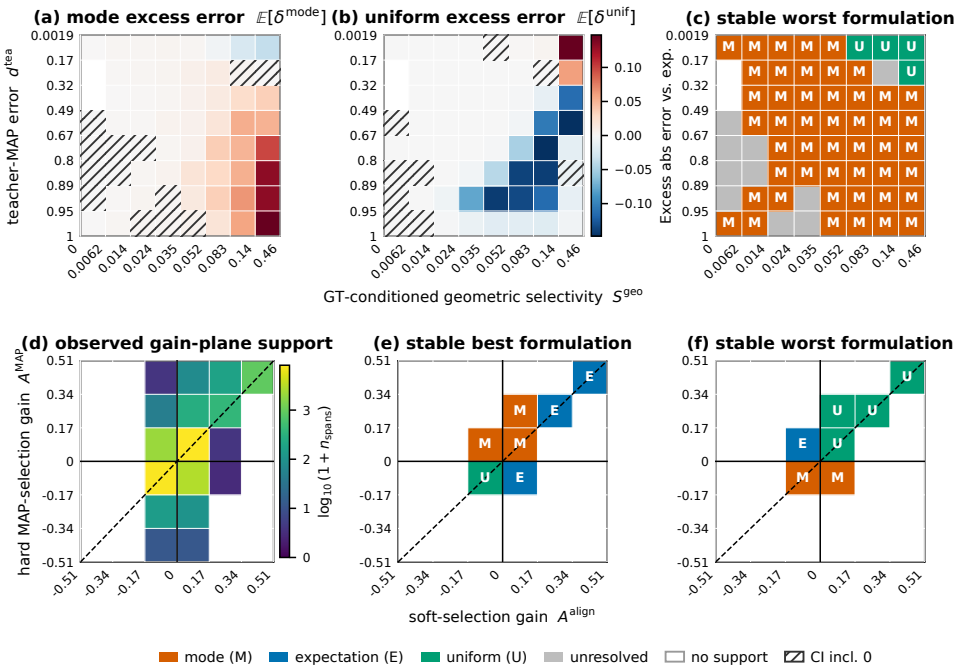

Figure 8: Conditional aggregation regimes and gain-space interpretation. Panels **(a)**–**(c)** use a shared 8×8 equal-frequency grid over S_j^{geo} and d_j^{tea}, with smaller, more reliable MAP errors at the top: **(a)** and **(b)** report the within-cell means of δ_j^{mode} and δ_j^{unif}, respectively, while **(c)** reports the stable worst formulation by conditional-mean error. Panels **(d)**–**(f)** use a symmetric fixed-width 6×6 grid over $(A_j^{\text{align}}, A_j^{\text{MAP}})$ in the lowest-d_j^{stu} tercile: **(d)** shows support, while **(e)** and **(f)** report the stable best and worst formulations. The solid zero axes and dashed diagonal mark the pairwise equal-risk boundaries in the idealized common-reference limit $d_j^{\text{stu}} = 0$; all rank colors are instead determined from the empirical errors e_j^g. A rank is shown only when the same formulation is uniquely best or worst in at least 95% of $1{,}000$ image-clustered bootstrap replicates.

shows that the observed gains concentrate near the origin, while Panels **(e)**–**(f)** report the best and worst formulations determined from the empirical errors e_j^g. Among the seven rank-resolved cells, Panel **(e)** splits the stable-best cells evenly between mode and expectation (*three each*): mode occupies the region where a positive hard gain exceeds the soft gain, expectation where a positive soft aligned gain dominates, and uniform wins only where neither form of selection is beneficial. Conversely, Panel **(f)** identifies uniform as worst most often (*four cells*) where both gains are positive, because flattening discards both forms of useful selection; mode is worst in two cells where hard MAP selection is harmful, whereas expectation is worst only once, where its soft aligned gain falls below both the hard gain and the uniform baseline. These observationsbroadly follow the analytic partition.

♢ *Expectation provides robust conditional ranking across regimes without oracle switching.* Among the 62 supported cells in the broader regime map, Panel **(c)** resolves a stable worst formulation in 52: mode is worst in 48, uniform in 4, and expectation in none, with the remaining 10 cells bootstrap-unresolved. The pointwise and rank-based diagnostics capture complementary properties: uniform achieves a bootstrap-resolved lower pointwise error than expectation in 49 cells, while Panel **(f)** identifies expectation as worst in only one resolved gain-plane cell. Across the broader conditional regime map, expectation therefore exhibits the most stable rank by avoiding both the widespread failure regime of hard MAP selection and the narrower failure regime induced by complete probability flattening. Switching between mode and uniform would require the annotation-relative quantities d_j^{tea}, A_j^{align}, and A_j^{MAP}, which are unavailable during training. MixOPD thus adopts expectation as a regime-agnostic approximation that preserves both the proposal support and its graded teacher mass.

♢ *Qualitative failure modes of endpoint aggregation.* Figure 9 makes the complementary endpoint failures concrete. In Panel **(a)**, the teacher MAP is farther from the reference than the student box; collapsing the retained evidence to this erroneous MAP causes mode to overestimate the student

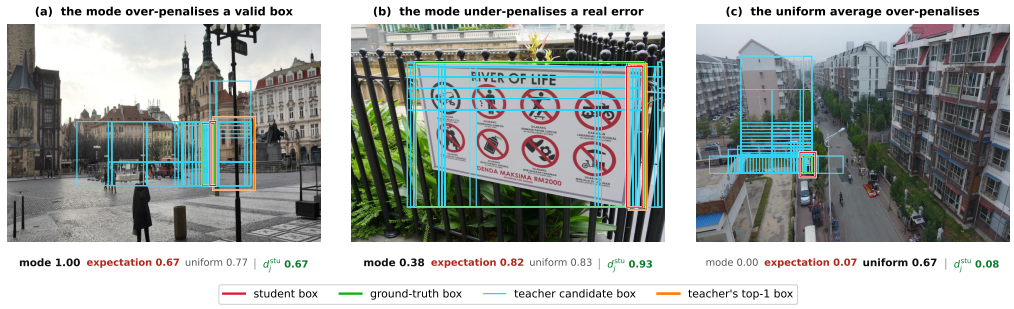

Figure 9: Representative failure cases of mode and uniform. Values report the three formulation estimates and the target d_j^{stu}. **(a)** An unreliable MAP causes mode to overestimate the error of a student box that is closer to the reference. **(b)** A similarly misplaced student–MAP pair causes mode to underestimate their annotation-relative error. **(c)** Equal weighting of dispersed tail candidates causes uniform to overestimate an otherwise small student error.

error (1.00 versus $d_j^{\mathrm{stu}} = 0.67$). In Panel **(b)**, the student and MAP share a similar localization error, so their mutual agreement masks the fact that both are far from the reference; mode consequently underestimates the error (0.38 versus 0.93). Thus, mode cannot distinguish correct agreement from shared error. Panel **(c)** shows the converse failure of uniform: although the student and MAP are accurate ($d_j^{\mathrm{stu}} = 0.08$), assigning equal mass to spatially dispersed tail candidates inflates the estimate to 0.67. Probability-weighted expectation avoids these endpoint commitments by retaining the proposal support while preserving its graded teacher mass.

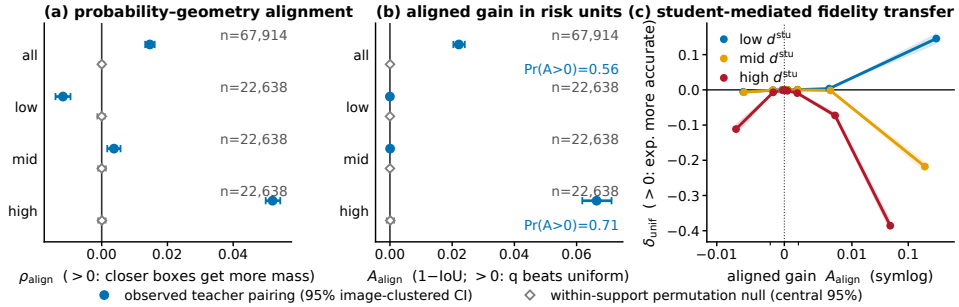

Figure 10: Teacher probability–geometry alignment and its transfer to formulation fidelity. **(a)** Mean alignment ρ_j^{align}, pooled and by joint-selection-capacity tercile, against a within-support permutation null. **(b)** The corresponding aligned gain A_j^{align} in $1 - \mathrm{IoU}$ risk units; positive values indicate that teacher weighting favors annotation-closer proposals over uniform weighting. **(c)** Uniform excess error δ_j^{unif} versus aligned gain, stratified by student error d_j^{stu}; positive values favor expectation.

Claim 2: Teacher probabilities provide a geometrically meaningful weighting of retained proposals. Teacher probability mass carries a weak but systematic signal of annotation-relative proposal quality. Figure 10**(a)** shows positive probability–geometry alignment overall ($\rho^{\mathrm{align}} = 0.0146$), increasing to 0.0520 in the highest-selection-capacity tercile and clearly separating from the within-support permutation null. Because the permutation preserves the proposal support and teacher probability values while breaking only their correspondence with proposal geometry, this separation identifies a non-random tendency to place greater mass on annotation-closer proposals. This ordering also has a measurable geometric consequence: Panel **(b)** reports a mean aligned gain of 0.0222 in $1 - \mathrm{IoU}$ risk units, rising to 0.0664 in the highest-capacity tercile, where 71% of spans have positive gain. Together, these results show that probability-weighted expectation converts the teacher's proposal ordering into a nontrivial, average reduction in annotation-relative candidate risk. Whether this teacher-side gain transfers to formulation fidelity depends on the student state. Among positive-gain spans in the low-d_j^{stu} regime, Panel **(c)** shows a mean uniform excess error of 0.0248, favoring expectation, whereas the relationship reverses in the middle- and high-error regimes. As training increases student IoU (Figure 6), realized spans increasingly move toward smaller d_j^{stu}, where positive teacher-side alignment is most likely to yield more faithful expectation-based calibration.

A.3 Fidelity and Efficiency of Teacher-Forced Constrained Coordinate Search

Formulation. TFCS differs from autoregressive beam search in both candidate scoring and support construction. We therefore probe the approximation in two stages: first by replacing its factorized scores with candidate-specific autoregressive scores on the identical retained support, and then by comparing the complete procedure with constrained beam search on its own support. The primary quantity is the scalar spatial coefficient $\widehat{\omega}_{j,K}$ that MixOPD actually applies during training.

Probing Setup. We construct an offline cohort of $1{,}000$ well-formed grounding spans from $1{,}000$ distinct images, balanced between images with one and multiple annotated regions. We use Qwen3-VL-8B-Instruct as the student and the same privileged Qwen3-VL-32B-Instruct teacher used for training. We use the training configuration $V = 64$, $M_{\mathrm{c}} = 64$, $K = 128$, and $s_{\min} = 16$. All comparisons share the same privileged prefix, realized span template, digit-only scoring convention, and geometric admissibility constraints. We evaluate three variants: *TFCS* is Algorithm 1; *AR-Same* fixes the TFCS support but recomputes every candidate score autoregressively, conditioning each digit on the candidate's own preceding digits; and *Beam* performs width-128 constrained autoregressive search over its own support. One TFCS case falls back, leaving 999 paired spans. We report Lin's concordance correlation coefficient (CCC), mean and median absolute error, and mean signed difference. For the same-support comparison, we additionally measure total variation $\widehat{\zeta}_j := \mathrm{TV}(\widehat{q}_j^{\mathrm{TFCS}}, \widehat{q}_j^{\mathrm{AR}})$ between the two normalized box laws; for the end-to-end comparison, we measure exact support Jaccard. Confidence intervals use $2{,}000$ image-level bootstrap replicates.

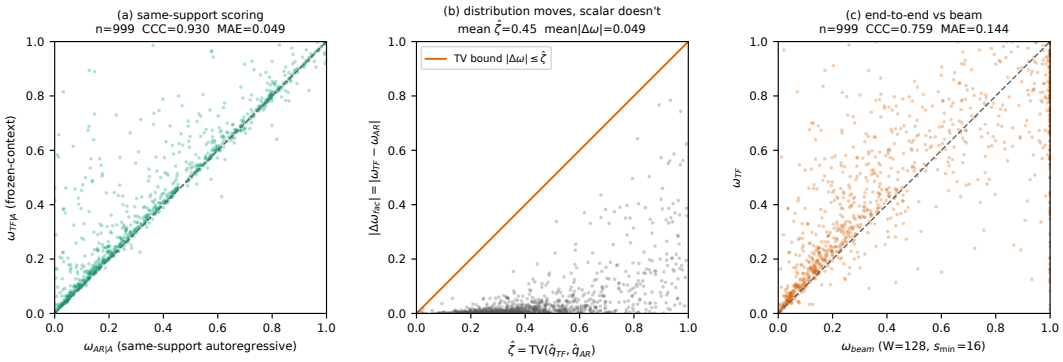

Figure 11: **Offline fidelity of teacher-forced constrained coordinate search.** (a) On the identical retained support, the spatial coefficients produced by factorized TFCS scoring and candidate-specific autoregressive rescoring closely agree. (b) The normalized box laws can differ substantially in total variation, while their induced spatial coefficients remain close; the solid line shows the bound $|\widehat{\omega}_j^{\mathrm{TFCS}} - \widehat{\omega}_j^{\mathrm{AR}}| \le \widehat{\zeta}_j$. (c) End-to-end comparison between TFCS and width-128 constrained beam search, each using its own retained support. Dashed diagonals denote exact agreement. Statistics are computed over the 999 successful paired spans.

Claim 1: TFCS preserves the spatial coefficient under autoregressive rescoring. As shown in Figure 11(a), on the identical candidate support, TFCS closely reproduces the coefficient obtained by candidate-specific autoregressive rescoring: CCC is 0.930, mean absolute error is 0.049 with 95% CI: $[0.043, 0.056]$, and median absolute error is only 0.011. Across the 999 spans, 78.6% have an absolute error of at most 0.05, and 93.6% remain within 0.20. More importantly, the normalized candidate laws differ by 0.453 in mean total variation, whereas the induced coefficients differ by only 0.049 on average, as presented in Figure 11(b). Consequently, frozen-context factorization can substantially redistribute probability among retained boxes while accurately preserving the scalar spatial divergence consumed by MixOPD.

Claim 2: TFCS retains substantial end-to-end agreement with beam search. As illustrated in Figure 11(c), when both scoring and support construction are allowed to differ, TFCS still tracks the spatial signal produced by width-128 beam search, attaining CCC 0.759, mean absolute error 0.144 with 95% CI: $[0.133, 0.155]$, and median absolute error 0.079. Their mean coefficients are 0.460 and 0.452, respectively, giving a mean difference of only $+0.008$ with a 95% confidence interval of $[-0.006, 0.022]$. This agreement holds even though their retained box sets have an exact Jaccard overlap of only 0.027. The much smaller same-support error (0.049 versus 0.144), together with

the low support overlap, indicates that candidate construction rather than factorized scoring is the primary source of the remaining discrepancy.

Table 6: Additional search cost beyond the rollout-level teacher-forcing pass. Rounds denote candidate-conditioned teacher decoding rounds. Beam totals use the observed mean and 90th-percentile numbers of grounding spans per rollout, $J = 4.55$ and $J = 14$, respectively.

Method	Rounds/span	Mean rollout	P90 rollout	Measured time/span
TFCS	0	0	0	~ 0.01 s
Beam ($W = 128$)	~ 15.5	~ 71	~ 217	~ 67 s

Claim 3: TFCS eliminates candidate-specific teacher decoding. TFCS reuses the single teacher-forcing pass already required by OPD and needs no additional teacher forwards for candidate construction. In contrast, beam search requires one prefill and an average of $L = 14.5$ sequential decoding rounds per grounding span, each operating over 128 beam states. Across 8,000 offline student rollouts, the number of grounding spans is $J = 4.55$ on average and $J = 14$ at the 90th percentile. Beam search therefore requires approximately 71 teacher decoding rounds for an average rollout and 217 at the 90th percentile, compared with zero additional rounds for TFCS, as reported in Table 6. Its bounded CPU enumeration takes approximately 0.01 seconds per span, compared with 67 seconds per span for beam search in our implementation. TFCS thereby reduces the measured span-wise search overhead by over three orders of magnitude and makes spatial calibration practical within online training.

Together, these results show that *TFCS preserves the scalar spatial signal required by MixOPD while avoiding the candidate-specific autoregressive computation of beam search.*

B Implementation Details

B.1 Training Data and Teacher Privileges



Figure 12: Illustration of annotation refinement for TreeVGR-RL (*left*) and ZwZ-RL-VQA (*right*).

We construct the training corpus from **TreeVGR-RL-37K** Wang et al. (2025a) and **ZwZ-RL-VQA** Wei et al. (2026), both of which target challenging visual understanding in high-resolution imagery. We first conduct a stringent leakage audit against all evaluation benchmarks, with particular attention to *V*Bench*, using dual-hash deduplication over image content and queries. This filtering yields approximately 145K samples. TreeVGR-RL-37K annotates each sample with boxes for its supporting visual evidence, whereas ZwZ-RL-VQA provides the image region supporting the ground-truth answer. We employ *DeepScan* Li et al. (2026b), a plug-and-play visually grounded reasoning agent instantiated with Qwen3-VL-8B-Instruct Bai et al. (2025a), to refine the retained evidence annotations from both datasets, as illustrated in Figure 12.

We then construct the two teacher privileges used throughout the paper. For the visual privilege, we highlight the refined evidence boxes in light green on the original image; for the logical privilege, we concatenate the question with its ground-truth answer (Figure 13). These inputs instantiate ξ_t^{V} and ξ_t^{T}, respectively, in the privileged teacher context $\widetilde{\mathbf{c}}_t = (\mathbf{c}_t, \xi_t^{\mathrm{V}}, \xi_t^{\mathrm{T}})$.

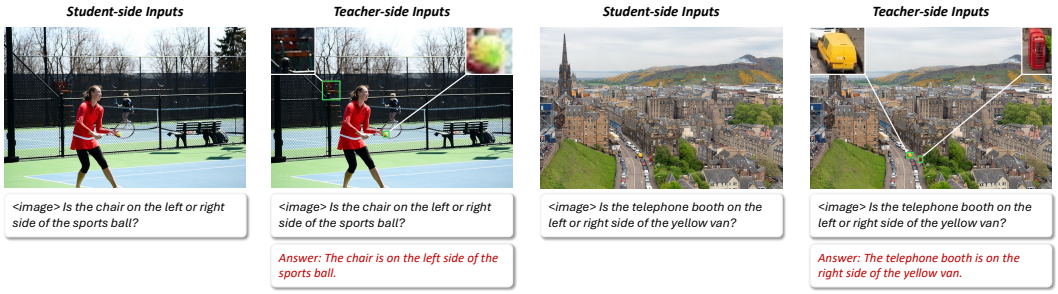

Figure 13: Student- and teacher-side inputs. The teacher receives the original image with the supporting evidence box highlighted in light green. The zoomed-in views are shown only for visualization and are not provided to the teacher.

B.2 Teacher-Forced Constrained Search

For each well-formed student grounding span b_j, we inherit its structural tokens and coordinate digit lengths $\{\ell_{j,c}\}_{c=1}^{4}$, and vary only the digit tokens. A single teacher-forcing pass over the realized student rollout provides

$$q_{j,c,r}(v) := \pi_{\text{tea}}\big(v \mid \widetilde{\mathbf{c}}_{t_{j,c,r}}\big)\,, \qquad \mathcal{D}_{j,c,r} := \mathcal{D}_{\text{dig}} \cap \text{TopV}(q_{j,c,r}, V)\,,$$

where $t_{j,c,r}$ is the position of digit r in coordinate c, and $\mathcal{D}_{\text{dig}}$ contains the ten single-digit tokens. Thus V refers to the full-vocabulary Top-V, after which nonnumeric tokens are discarded. A coordinate value $z = \text{int}(d_{1:\ell_{j,c}})$ receives score $\sum_r \log q_{j,c,r}(d_r)$. For a candidate box b, summing its four coordinate scores gives the factorized score

$$u_j(b) = \sum_{c=1}^{4} \sum_{r=1}^{\ell_{j,c}} \log q_{j,c,r}(d_{c,r}(b))\,.$$

At each digit position within a coordinate, we extend the current partial-value list with the retained digit tokens and immediately keep its Top-M_c entries. The four completed coordinate lists are then combined by tie-aware best-first Cartesian search. Before applying cutoff K, each axis-pair stage admits only pairs satisfying $z_2 - z_1 \geq s_{\min} = 16$, which also enforces the natural ordering; out-of-range pairs are subsequently removed. The final box merge also uses cutoff K, retains the complete boundary tie group, and applies a fixed final tie-break to select at most K boxes. On success, the resulting set is denoted by $\mathcal{B}^K_{\text{tea},j} := \mathcal{B}^K_{\text{tea}}(\widetilde{\mathbf{c}}_{a_j})$. Normalizing u_j over this set gives $\widehat{q}_j(b) := \widehat{P}^{\mathcal{B}}_{\text{tea},K}(b \mid \widetilde{\mathbf{c}}_{a_j})$ in Eq. 2. Malformed or unparsed spans bypass the search and retain unit weight. If digit alignment or candidate construction fails, including when either constrained axis list is empty, the search returns $(\varnothing, \varnothing, 1)$, recovering the base surrogate. Intuitively, position-wise truncation prevents exponential growth within each coordinate, while the staged Cartesian merges efficiently assemble the retained box set without candidate-specific decoding. The axis-level admissibility predicates remove sliver boxes before either axis cutoff, ensuring that every returned box satisfies the minimum-side constraint.

B.3 Training Objectives and Recipe

All variants are initialized directly from Qwen3-VL-4B/8B-Instruct Bai et al. (2025a), without a separate cold-start stage. We train with VeRL Sheng et al. (2024) and FSDP, using vLLM 0.11.0 Kwon et al. (2023) for rollout generation. For all OPD-based variants, we use a frozen Qwen3-VL-32B-Instruct teacher (Appendix D.1), additionally conditioned on the visual and logical privileges described in Appendix B.1. Both teacher and student use the visually grounded reasoning prompt in Appendix B.4. Because the privileged inputs alter the teacher's multimodal prefix, we independently construct its input IDs, position IDs, and attention masks at every step rather than reusing the student-side tensors. Student rollouts are always sampled from $\pi_{\theta_{\text{old}}}$ under the unprivileged context $\mathbf{c}_t$; the teacher only scores these rollouts under $\widetilde{\mathbf{c}}_t$.

GRPO. For each response $\mathbf{x}_i$, a frozen Qwen3-VL-32B-Instruct judge conditioned on the ground-truth answer assigns a binary accuracy reward $R_{\mathrm{acc},i} \in \{0,1\}$. We extract all valid predicted boxes and compute their *dual IoU* Li et al. (2026a) against the ground-truth annotations, yielding $R_{\mathrm{dIoU},i} \in [0,1]$. Given a group of G responses, we define

$$R_i = R_{\mathrm{acc},i} + R_{\mathrm{dIoU},i}, \qquad \widehat{A}_i = \mathrm{sg}\left[\frac{R_i - \mathrm{mean}_{j\in[G]}\, R_j}{\mathrm{std}_{j\in[G]}\, R_j + \epsilon_{\mathrm{adv}}}\right], \qquad \varrho_{i,t}(\theta) = \frac{\pi_\theta(x_{i,t} \mid \mathbf{c}_{i,t})}{\pi_{\theta_{\mathrm{old}}}(x_{i,t} \mid \mathbf{c}_{i,t})}.$$

For compactness, define the asymmetrically clipped term

$$\ell_{i,t}^{\mathrm{clip}}(\theta;\epsilon_-,\epsilon_+) = \max\Big\{-\varrho_{i,t}(\theta)\widehat{A}_i, -\mathrm{clip}(\varrho_{i,t}(\theta), 1-\epsilon_-, 1+\epsilon_+)\,\widehat{A}_i\Big\}.$$

VeRL additionally applies dual clipping at negative-advantage positions:

$$\ell_{i,t}^{(c)}(\theta;\epsilon_-,\epsilon_+) = \begin{cases} \min\Big\{\ell_{i,t}^{\mathrm{clip}}(\theta;\epsilon_-,\epsilon_+), -c\widehat{A}_i\Big\}, & \widehat{A}_i < 0, \\ \ell_{i,t}^{\mathrm{clip}}(\theta;\epsilon_-,\epsilon_+), & \widehat{A}_i \geq 0. \end{cases}$$

GRPO Guo et al. (2025) shares $\widehat{A}_i$ across all tokens of response i and reduces both its symmetrically clipped policy term and KL regularizer by a single mean over all valid response tokens. Here and below, the expectation is over $\mathbf{q} \sim \mathcal{Q}$ and $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \pi_{\theta_{\mathrm{old}}}(\cdot \mid \mathbf{q})$ for $i \in [G]$:

$$\mathcal{L}_{\mathrm{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{\sum_{i=1}^{G} T_i}\sum_{i=1}^{G}\sum_{t=1}^{T_i}\left[\ell_{i,t}^{(C_{\mathrm{GRPO}})}(\theta;\epsilon_{\mathrm{clip}},\epsilon_{\mathrm{clip}}) + \beta\widehat{\mathcal{K}}_{i,t}^{\mathrm{ref}}(\theta)\right]\right].$$

Here $\widehat{\mathcal{K}}_{i,t}^{\mathrm{ref}}$ is VeRL's `low_var_kl` (K3) token-level estimator, and the reference policy is a frozen copy of the corresponding Qwen3-VL-4B/8B-Instruct initialization.

DAPO. Yu et al. (2026a) uses the reward configuration and group-normalized response advantage. It removes reference-policy KL regularization, adopts asymmetric Clip-Higher with $\epsilon_{\mathrm{high}} > \epsilon_{\mathrm{low}}$, and retains VeRL's dual clip with a method-specific threshold C_{DAPO}. Its policy loss is

$$\mathcal{L}_{\mathrm{DAPO}}(\theta) = \mathbb{E}\left[\frac{1}{\sum_{i=1}^{G} T_i}\sum_{i=1}^{G}\sum_{t=1}^{T_i}\ell_{i,t}^{(C_{\mathrm{DAPO}})}(\theta;\epsilon_{\mathrm{low}},\epsilon_{\mathrm{high}})\right].$$

Following VeRL's vision-language DAPO recipe, we enable dynamic group filtering on accuracy: each generation batch contains three times as many prompt groups as a policy-update batch, groups with uniformly correct or uniformly incorrect responses are discarded, and sampling continues for at most four generation batches. Overlong reward shaping is disabled.

Standard OPD. We follow the OPD recipe Lu & Lab (2025). Its sampled reverse-KL signal is detached and used as a dense token-level advantage in a policy-gradient update, rather than directly backpropagated as an analytic KL loss. Let

$$A_t^{\mathrm{raw}} = \log\frac{\pi_{\mathrm{tea}}(x_t \mid \widetilde{\mathbf{c}}_t)}{\pi_{\theta_{\mathrm{old}}}(x_t \mid \mathbf{c}_t)}, \qquad \widehat{A}_t = \mathrm{sg}[\mathrm{clip}(A_t^{\mathrm{raw}}, -C_A, C_A)], \qquad \varrho_t(\theta) = \frac{\pi_\theta(x_t \mid \mathbf{c}_t)}{\pi_{\theta_{\mathrm{old}}}(x_t \mid \mathbf{c}_t)}.$$

The implemented PPO-clipped local surrogate is

$$\ell_t^{\mathrm{OPD}}(\theta) = \max\Big\{-\varrho_t(\theta)\widehat{A}_t, -\mathrm{clip}(\varrho_t(\theta), 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}})\,\widehat{A}_t\Big\}.$$

For a rollout minibatch $\mathcal{M}$ with valid-response-token mask m_t, standard OPD minimizes

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{\mathcal{M}}\left[\frac{\sum_{\mathcal{M}} m_t \ell_t^{\mathrm{OPD}}(\theta)}{\sum_{\mathcal{M}} m_t}\right].$$

Entropy-aware OPD. Following EOPD Jin et al. (2026), we retain ℓ_t^{OPD} at every valid response position and additionally apply forward KL at high-entropy teacher positions. Let

$$q_t = \pi_{\mathrm{tea}}(\cdot \mid \widetilde{\mathbf{c}}_t), \qquad g_t = \mathbf{1}\{H(q_t) > \eta_H\},$$

and let $\bar{q}_t^{\,L}$ denote q_t renormalized over its Top-L support $\mathcal{S}_t^L$. We define

$$F_t(\theta) = D_{\mathrm{KL}}\big(\bar{q}_t^{\,L} \big\| \pi_\theta(\cdot \mid \mathbf{c}_t)\big).$$

The implemented local surrogate is $S_t^{(\lambda_F)}(\theta) = \ell_t^{\mathrm{OPD}}(\theta) + \lambda_F g_t F_t(\theta)$, and the uncalibrated additive-EOPD objective is

$$\mathcal{L}_{\mathrm{EOPD}}(\theta) = \mathbb{E}_{\mathcal{M}}\left[\frac{\sum_{\mathcal{M}} m_t S_t^{(\lambda_F)}(\theta)}{\sum_{\mathcal{M}} m_t}\right].$$

Thus $\lambda_F = 0$ recovers standard OPD, whereas $\lambda_F > 0$ gives EOPD.

MixOPD. For each parsed, well-formed grounding span j, MixOPD applies the teacher-forced constrained coordinate search in Appendix B.2. From the teacher distributions cached along the realized student rollout, the search returns the retained boxes $\mathcal{B}_{\mathrm{tea},j}^K$ and their normalized weights $\widehat{q}_j$, as defined there. The corresponding spatial coefficient is

$$\widehat{\omega}_{j,K} = \sum_{b \in \mathcal{B}_{\mathrm{tea},j}^K} \widehat{q}_j(b)\,[1 - \mathrm{IoU}(b_j, b)]\,.$$

If candidate construction fails, we set $\widehat{\omega}_{j,K} = 1$, recovering the base local surrogate. For the pairwise-disjoint parsed grounding spans, define

$$w_t = \begin{cases} \mathrm{sg}(\widehat{\omega}_{j,K}), & \exists_{=1}\, j :\ t \in \mathcal{I}(b_j), \\ 1, & \text{otherwise.} \end{cases}$$

Hence reasoning, answer, structural, and malformed-span positions retain unit weight. MixOPD calibrates the complete implemented local surrogate:

$$\mathcal{L}_{\mathrm{MixOPD}}^{(\lambda_F)}(\theta) = \mathbb{E}_{\mathcal{M}}\left[\frac{\sum_{\mathcal{M}} m_t w_t S_t^{(\lambda_F)}(\theta)}{\sum_{\mathcal{M}} m_t}\right].$$

The denominator is the number of valid response tokens, rather than a span length or a sum of spatial weights. Appendix B.2 specifies candidate construction and fallback behavior; complete training pseudocode and complexity are given in Appendix B.6.

All rollout-fixed and teacher-side quantities, together with outputs of the parser, entropy gate, coordinate search, and geometry computations, are detached; gradients flow only through the live policy ratio $\varrho_t(\theta)$ and, for EOPD, the forward-KL term $F_t(\theta)$. At inference, MixOPD uses the student alone, with no teacher, search, additional parameters, or extra model pass.

B.4 Hyperparameters

Unless otherwise specified, all methods share the same optimization recipe. We optimize the student with AdamW Loshchilov & Hutter (2017), using $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 0.01, and gradient-norm clipping at 1.0. The learning rate is linearly warmed up to 1×10^{-6} over the first 3% of optimizer steps and held constant thereafter. Each policy-update batch contains 256 prompt groups and is split into PPO minibatches of 32 groups. We sample one response per prompt for OPD, EOPD, and MixOPD, and eight responses per prompt for GRPO and DAPO. Both RL baselines and all OPD-based objectives use a global mean over valid response tokens. DAPO generates 768 prompt groups per sampling batch, dynamically retains 256 nondegenerate groups according to binary accuracy, and allows at most four generation batches to fill an update batch.

For both RL baselines, VeRL's GRPO estimator normalizes group rewards by their standard deviation with $\epsilon_{\mathrm{adv}} = 10^{-6}$, and the entropy-bonus coefficient is zero. GRPO uses $\epsilon_{\mathrm{clip}} = 0.2$, dual-clip threshold $C_{\mathrm{GRPO}} = 3.0$, and a `low_var_kl` reference loss with $\beta = 0.01$, following VeRL's Qwen3-VL configuration. DAPO uses $(\epsilon_{\mathrm{low}}, \epsilon_{\mathrm{high}}) = (0.2, 0.28)$, $C_{\mathrm{DAPO}} = 10.0$, and no KL term in either the reward or actor loss. Its overlong reward shaping is disabled.

For OPD objectives, the sampled teacher–student log-ratio is clipped with radius $C_A = 10$, and the PPO likelihood ratio uses $\epsilon_{\mathrm{clip}} = 0.2$. For EOPD variants, we set $\lambda_F = 1.0$ and apply the forward-KL term when the full-vocabulary teacher entropy exceeds $\eta_H = 1.2$ nats. The teacher target is renormalized over its Top-L support with $L = 64$. The teacher-forced constrained coordinate search in Appendix B.2 uses $V = 64$, $M_c = 64$, $K = 128$, and $s_{\min} = 16$. Although $L = V = 64$ in our configuration, L specifies the EOPD target support, whereas V specifies the full-vocabulary teacher shortlist at each realized digit position. The cap M_c is applied to the partial coordinate-value

list after every digit-position expansion, whereas K sets the cutoff for both axis pairing and the final box merge. Each retained rollout batch is traversed once for policy updates (`ppo_epochs`=1), and all student policies are trained for one pass over the training set on 16× NVIDIA H800 80GB GPUs. DAPO may generate additional candidate groups before forming each retained update batch. Table 7 summarizes the key configuration.

Table 7: Training hyperparameters for MixOPD and the reproduced baselines. Method-specific settings are indicated explicitly.

Hyperparameter	**Value**	**Hyperparameter**	**Value**
Student policy	Qwen3-VL-4B/8B	Teacher (OPD-based)	Qwen3-VL-32B
Training framework	VeRL/FSDP	Rollout engine	vLLM 0.11.0
Verifiable reward (GRPO/DAPO)	Acc. + dual IoU	Optimizer	AdamW
β_1, β_2	0.9, 0.999	Weight decay	0.01
Peak learning rate	1×10^{-6}	LR schedule	3% warm-up; constant
Global prompt batch	256	PPO minibatch	32 prompt groups
PPO clip ϵ_{clip} (OPD-based)	0.2	Policy-update epochs	1
Rollouts / prompt (OPD-based)	1	Rollouts / prompt (GRPO/DAPO)	8
Sampling temperature	1.0	Sampling top-p	1.0
Max prompt length	11,264	Max response length	1,024
Advantage ϵ_{adv}	1×10^{-6}	RL loss aggregation	Valid-token mean
GRPO clip $(\epsilon_-, \epsilon_+, c)$	$(0.2, 0.2, 3)$	DAPO clip $(\epsilon_-, \epsilon_+, c)$	$(0.2, 0.28, 10)$
GRPO KL loss	$\beta = 0.01$; K3	DAPO KL (reward/loss)	Disabled
GRPO reference	Frozen initial student	DAPO dynamic sampling	3× gen.; acc.; max 4
RL entropy bonus	0	DAPO overlong shaping	Disabled
Sampled-log-ratio radius C_A	10	Entropy threshold η_H	1.2 nats
Additive FKL weight λ_{F}	1.0	FKL teacher support L	64
Search Top-V logits	64	Position-wise partial-value cap M_{c}	64
Minimum side $s_{\min}$	16	Maximum retained boxes K	128
Search merge	Tie-aware best-first; deterministic final tie-break	Rollout tensor parallelism	8
Actor sequence parallelism	2	Gradient clipping	1.0
Training epochs	1	Hardware	16× H800 80GB

B.5 Prompt

All variants (GRPO, DAPO, OPD, EOPD, and MixOPD counterparts) use the *same* system prompt:

> "*You are a helpful multimodal assistant. Given an image and a user question, respond with exactly two blocks in the following order. First, output a single* `<thinking>...</thinking>` *block containing two stages: analyze and interpret the question, determining what must be recognized, located, counted, compared, read, or inferred from the image; and then identify the image regions that most directly support the answer. In the second stage, use an interleaved box–explanation format: for each region, first output its bounding box exactly as* `<box>`x_1, y_1, x_2, y_2`</box>`*, where* (x_1, y_1) *and* (x_2, y_2) *denote the top-left and bottom-right corners, respectively, and the integer coordinates satisfy* $0 \le x_1 < x_2 \le 1000$ *and* $0 \le y_1 < y_2 \le 1000$*; then immediately explain how that region helps answer the question before introducing another box. Second, output a single* `<answer>...</answer>` *block containing a clear, concise answer based only on the visual evidence established within the* `<thinking>` *block. Output no text outside these two blocks.*"

B.6 Complete Training Algorithms and Complexity

Algorithms 2 and 3 instantiate MixOPD with standard OPD and EOPD, respectively. Both use the clipped sampled teacher–student log ratio, which corresponds to the negative sampled K_1 reverse-KL estimator, as a detached token advantage inside a PPO-clipped surrogate. EOPD additionally applies an entropy-gated Top-L forward-KL term. The rollout, parser, teacher-forced coordinate search (Algorithm 1), and spatial weights are otherwise shared. The complete update sequence is given in the two algorithms below.

End-to-end OPD computation. We analyze one retained global rollout–update cycle with batch size B. For example i, let S_i, $\widetilde{S}_i$, and T_i denote the student-context, privileged-teacher-context, and

response lengths, and define

$$N_{\mathrm{s}} \coloneqq \sum_{i=1}^{B}(S_i + T_i), \qquad N_{\mathrm{t}} \coloneqq \sum_{i=1}^{B}(\widetilde{S}_i + T_i), \qquad N_{\mathrm{r}} \coloneqq \sum_{i=1}^{B} T_i.$$

Let P_{s} and P_{t} be the student and teacher parameter counts, E_{PPO} the number of policy-update epochs, and N_{mb} the number of policy minibatches per epoch. Using the standard dense-Transformer parameter-multiplication estimate of approximately $2PN$ FLOPs for a forward pass and $6PN$ for a forward–backward pass, one cycle costs

$$\mathcal{F}_{\mathrm{OPD}} \approx 2P_{\mathrm{s}}N_{\mathrm{s}} + 2P_{\mathrm{t}}N_{\mathrm{t}} + 6E_{\mathrm{PPO}}P_{\mathrm{s}}N_{\mathrm{s}} + O((1 + E_{\mathrm{PPO}})N_{\mathrm{r}} + E_{\mathrm{PPO}}N_{\mathrm{mb}}P_{\mathrm{s}})\,. \tag{5}$$

The first three terms are student rollout generation, frozen-teacher forcing, and the live-student forward–backward update. The last term contains the one-time tokenwise estimator, per-epoch PPO arithmetic, and the parameterwise optimizer updates performed once per policy minibatch. In particular,

$$\widehat{k}_{1,t} \coloneqq \log p_t^{\mathrm{old}}(x_t) - \log q_t(x_t), \qquad \widehat{A}_t = \mathrm{sg}\Big[\mathrm{clip}(-\widehat{k}_{1,t}, -C_A, C_A)\Big]\,.$$

Once the sampled log probabilities are available, constructing $\widehat{k}_{1,t}$ and $\widehat{A}_t$ takes $\Theta(N_{\mathrm{r}})$ scalar operations. Forming the PPO ratio, clipping, and reducing the loss take $\Theta(N_{\mathrm{r}})$ per policy epoch, while an AdamW-style optimizer touches $\Theta(P_{\mathrm{s}})$ parameter states per minibatch. OPD never forms a vocabulary-wide analytic reverse KL. For our $B = 256$, $E_{\mathrm{PPO}} = 1$, $P_{\mathrm{s}} = 8\times10^9$, and $P_{\mathrm{t}} = 32\times10^9$, Equation 5 becomes

$$\mathcal{F}_{\mathrm{OPD}} \approx 64 \times 10^9 (N_{\mathrm{s}} + N_{\mathrm{t}}) \text{ FLOPs}.$$

If the two sides have the same average full-sequence length $\bar{N}$, this is

$$\mathcal{F}_{\mathrm{OPD}} \approx 32.768 \left(\frac{\bar{N}}{10^3}\right) \text{PFLOPs}$$

per retained global cycle. The table reports length limits rather than measured means. Under the maximum-length envelope

$$N_{\max} = 256(11{,}264 + 1{,}024) = 3{,}145{,}728,$$

and assuming both sides attain this cap, the parameter-multiplication estimate is

$$\underbrace{50.33}_{\text{student rollout}} + \underbrace{201.33}_{\text{teacher forcing}} + \underbrace{150.99}_{\text{student forward–backward}} = 402.65 \text{ PFLOPs}.$$

This is a reproducible maximum-length arithmetic envelope, not a measured typical step cost. The eight policy minibatches of size 32 partition the retained batch and therefore do not multiply its aggregate model forward–backward FLOPs when $E_{\mathrm{PPO}} = 1$; they do induce eight parameterwise optimizer sweeps, captured by the lower-order remainder in Equation 5. Non-parameter-matrix attention terms, visual encoding, optimizer-state traffic, distributed communication, and kernel overhead are omitted from this conventional estimate; all are already present in OPD and increase the baseline against which MixOPD is compared.

Additional EOPD computation. Let $U \coloneqq |\mathcal{V}|$ and $N_H \coloneqq \sum_t m_t g_t \le N_{\mathrm{r}}$. EOPD reuses the teacher and live-student logits produced by the passes above. Teacher entropy and Top-L selection require $O(N_{\mathrm{r}}U)$ cached-logit work with a standard linear-scan selection primitive, while the gated, truncated forward KL costs $O(E_{\mathrm{PPO}}N_H L)$, bounded by $O(E_{\mathrm{PPO}}N_{\mathrm{r}}L)$. With $E_{\mathrm{PPO}} = 1$, $L = 64$, and $N_{\mathrm{r}} \le 256 \times 1{,}024 = 262{,}144$, the latter evaluates at most

$$N_{\mathrm{r}}L = 16{,}777{,}216$$

support terms per retained cycle. These are elementwise or selection operations on available logits and require no additional model pass.

Additional MixOPD computation. Let $J_{\mathrm{tot}} \coloneqq \sum_{i=1}^{B} J_i$ be the total number of parsed grounding spans, let $L_j = |\mathcal{I}(b_j)|$, and define

$$L_{\mathrm{coord}} \coloneqq \sum_{j=1}^{J_{\mathrm{tot}}} L_j \le N_{\mathrm{r}}.$$

Top-V selection at these cached teacher-logit rows costs $O(L_{\mathrm{coord}}U)$ with a linear-scan selection primitive. For coordinate h, write

$$d_{j,h,r} \coloneqq |\mathcal{D}_{j,h,r}| \le d_{\max} \coloneqq \min\{V, 10\}.$$

The position-wise partial-value truncation satisfies

$$E_{j,h,r} = n_{j,h,r-1} d_{j,h,r}, \qquad n_{j,h,r} = \min\{M_{\mathrm{c}}, E_{j,h,r}\}, \qquad n_{j,h,0} = 1.$$

Thus $E_{j,h,r} \le M_{\mathrm{c}} d_{\max}$, and coordinate enumeration and sorting cost

$$\mathcal{C}_{\mathrm{enum}} = O\left(\sum_{j=1}^{J_{\mathrm{tot}}} \sum_{h=1}^{4} \sum_{r=1}^{\ell_{j,h}} E_{j,h,r} \log(1 + E_{j,h,r}) \right). \tag{6}$$

Let $H_{j,x}$, $H_{j,y}$, and $H_{j,b}$ be the states examined by the two ordered axis merges and the final box merge, counting each Cartesian state at most once. In particular, $H_{j,x}$ and $H_{j,y}$ count all examined states, including those rejected by the minimum-side predicate and those inspected to complete the cutoff-tie group. Their heap and final tie-sorting cost is

$$\mathcal{C}_{\mathrm{merge}} = O\left(\sum_{j=1}^{J_{\mathrm{tot}}} \sum_{a \in \{x,y,b\}} H_{j,a} \log(1 + H_{j,a}) \right). \tag{7}$$

If each constrained merge completes its cutoff after examining $O(K)$ states, this reduces to $O(J_{\mathrm{tot}} K \log K)$. If $R_{j,x}$ and $R_{j,y}$ are the returned axis-list sizes, the finite ceilings are

$$H_{j,x}, H_{j,y} \le M_{\mathrm{c}}^2, \qquad H_{j,b} \le R_{j,x} R_{j,y} \le M_{\mathrm{c}}^4.$$

On a successful search, for $1 \le K_j = |\mathcal{B}^K_{\mathrm{tea},j}| \le K$, the spatial coefficient is

$$\widehat{\omega}_{j,K} = \sum_{b \in \mathcal{B}^K_{\mathrm{tea},j}} \widehat{q}_j(b) \left[1 - \mathrm{IoU}(b_j, b)\right] = \mathcal{D}_{\mathrm{spa}}(\delta_{b_j}, \widehat{q}_j).$$

Normalization, K_j IoUs, and the weighted reduction take $\Theta(K_j)$ time; broadcasting the scalar takes $\Theta(L_j)$. Hence

$$\mathcal{C}_{\mathrm{spa}} = \Theta\left(\sum_{j=1}^{J_{\mathrm{tot}}} K_j + L_{\mathrm{coord}} \right) = O(J_{\mathrm{tot}} K + N_{\mathrm{r}}). \tag{8}$$

On search fallback, $K_j = 0$, $\widehat{q}_j = \varnothing$, and $\widehat{\omega}_{j,K} = 1$; the preceding expectation formula does not apply, and normalization and IoU evaluation are skipped. The constant-time minimum-side checks occur while examining axis-merge states and are already included in $H_{j,x}$, $H_{j,y}$, and $\mathcal{C}_{\mathrm{merge}}$; there is no post-retention minimum-side scan. The final weighted-token reduction is part of the per-epoch $O(E_{\mathrm{PPO}} N_{\mathrm{r}})$ remainder in Equation 5, rather than $\mathcal{C}_{\mathrm{spa}}$. Sequential span processing uses $O(K)$ transient storage for spatial calibration; search storage is output-sensitive in the states of Equation 7.

Overall order and configured scale. The two variants satisfy

$$\mathcal{C}^{\mathrm{std}}_{\mathrm{MixOPD}} = \mathcal{F}_{\mathrm{OPD}} + O(L_{\mathrm{coord}} U) + \mathcal{C}_{\mathrm{enum}} + \mathcal{C}_{\mathrm{merge}} + \mathcal{C}_{\mathrm{spa}}, \tag{9}$$

$$\mathcal{C}^{\mathrm{EOPD}}_{\mathrm{MixOPD}} = \mathcal{C}^{\mathrm{std}}_{\mathrm{MixOPD}} + O(N_{\mathrm{r}} U + E_{\mathrm{PPO}} N_H L). \tag{10}$$

For the fixed training configuration $(L, V, M_{\mathrm{c}}, K) = (64, 64, 64, 128)$, bounded coordinate length, and fixed model vocabulary, pairwise-disjoint spans imply $L_{\mathrm{coord}}, J_{\mathrm{tot}} = O(N_{\mathrm{r}})$. Consequently, both variants retain the leading model-compute order

$$\mathcal{C}^{\mathrm{std}}_{\mathrm{MixOPD}}, \mathcal{C}^{\mathrm{EOPD}}_{\mathrm{MixOPD}} = \Theta(P_{\mathrm{s}} N_{\mathrm{s}} + P_{\mathrm{t}} N_{\mathrm{t}}).$$

The search hyperparameters occur only in the explicit additive terms above; they neither enter the leading neural-compute term nor change the number of model evaluations. The configured constants also make the scale explicit. With $V = M_c = 64$, $K = 128$, and at most four digits per coordinate, one coordinate expands at most

$$10 + 100 + 640 + 640 = 1{,}390$$

partial values, or $5{,}560$ across a box. Sorting the four lists has the comparison-count proxy

$$4[10 \log_2 10 + 100 \log_2 100 + 2(640 \log_2 640)] \approx 5.05 \times 10^4,$$

and the all-admissible, cutoff-only reference for the three merges is

$$3K \log_2 K = 2{,}688.$$

Spatial normalization, IoU evaluation, reduction, and broadcasting add at most $3K + 16 = 400$ scalar visits or assignments. Under this $H_{j,a} = O(K)$ reference and omitting the constant-time predicate checks already counted by $H_{j,x} + H_{j,y}$, the proxy excluding the cached-logit Top-V scan is approximately 5.36×10^4 list, heap, and reduction primitives per span, or $1.37 \times 10^7 \bar{J}$ for 256 rollouts with mean span count $\bar{J}$. Actual merge work is governed by Equation 7. The separately accounted Top-V selection scans at most $16U$ cached-logit entries per span. These CPU-oriented list and heap counts are not interchangeable with accelerator FLOPs; they expose the scale separation from the model passes.

Measured runtime. The post-forward candidate construction and spatial calibration take approximately 0.01 s of CPU time per span in our implementation, or approximately $0.01\bar{J}$ s per rollout. For a one-span low-resolution rollout, this is about 1.7% of the corresponding 0.6 s 32B-teacher forward alone; the fraction is smaller against the complete OPD cycle and for high-resolution inputs, whose teacher forward takes tens of seconds. These operations are used only during training.

ALGORITHMIC DETAILS

Algorithm 1 Teacher-Forced Constrained Coordinate Search

Require: Parsed span b_j; cached teacher laws $\{q_{j,c,r}\}$; coordinate lengths $\{\ell_{j,c}\}$; V, M_c, K; $s_{\min} = 16$
Ensure: $\mathcal{B}^K_{\text{tea},j}$, $\widehat{q}_j$, and $\widehat{\omega}_{j,K}$

FAIL $\leftarrow (\varnothing, \varnothing, 1)$; **if** digit alignment fails, **return** FAIL
for $c = 1, \ldots, 4$ **do**
  $(\mathbf{z}, \mathbf{u}) \leftarrow ([0], [0])$
  for $r = 1, \ldots, \ell_{j,c}$ **do**
    $I \leftarrow [i \in \text{STABLETOPKIDX}(q_{j,c,r}, V) : \text{isDigit}(i)]$; **if** $I = \varnothing$, **return** FAIL
    $\mathbf{d} \leftarrow \text{digit}(I)$, $\boldsymbol{\lambda} \leftarrow \log q_{j,c,r}[I]$
    $\mathbf{z}' \leftarrow \text{vec}(10\mathbf{z}[:, \text{None}] + \mathbf{d}[\text{None}, :])$, $\mathbf{u}' \leftarrow \text{vec}(\mathbf{u}[:, \text{None}] + \boldsymbol{\lambda}[\text{None}, :])$
    $J \leftarrow \text{STABLETOPKIDX}(\mathbf{u}', M_c)$; $(\mathbf{z}, \mathbf{u}) \leftarrow (\mathbf{z}'[J], \mathbf{u}'[J])$
  $\mathcal{Z}_{j,c} \leftarrow (\mathbf{z}, \mathbf{u})$
$\widetilde{\mathcal{X}}_j \leftarrow \text{PAIRTOPTIE}(\mathcal{Z}_{j,1}, \mathcal{Z}_{j,3}, K, \lambda(a, b) : b.z - a.z \geq s_{\min})$
$\widetilde{\mathcal{Y}}_j \leftarrow \text{PAIRTOPTIE}(\mathcal{Z}_{j,2}, \mathcal{Z}_{j,4}, K, \lambda(a, b) : b.z - a.z \geq s_{\min})$
$\mathcal{X}_j \leftarrow \widetilde{\mathcal{X}}_j \cap [0, 1000]^2$, $\mathcal{Y}_j \leftarrow \widetilde{\mathcal{Y}}_j \cap [0, 1000]^2$; **if** $\mathcal{X}_j = \varnothing$ or $\mathcal{Y}_j = \varnothing$, **return** FAIL
$\widetilde{\mathcal{C}}_j \leftarrow \text{PAIRTOPTIE}(\mathcal{X}_j, \mathcal{Y}_j, K, \lambda(a, b) : \text{True})$
$\mathcal{C}_j \leftarrow \text{sorted}(\widetilde{\mathcal{C}}_j; -u_b, i_{x_1}, i_{y_1}, i_{x_2}, i_{y_2})[:K]$
$\mathcal{B}^K_{\text{tea},j} \leftarrow [b : (b, u_b) \in \mathcal{C}_j]$, $\widehat{\mathbf{q}}_j \leftarrow \text{softmax}([u_b : (b, u_b) \in \mathcal{C}_j])$
$\widehat{\omega}_{j,K} \leftarrow \sum_{b \in \mathcal{B}^K_{\text{tea},j}} \widehat{q}_j(b)\,[1 - \text{IoU}(b_j, b)]$
return $(\mathcal{B}^K_{\text{tea},j}, \widehat{q}_j, \widehat{\omega}_{j,K})$

STABLETOPKIDX returns at most k indices in descending-score order with a fixed tie break. PAIRTOPTIE performs the implementation's lazy best-first Cartesian merge. Its predicate is evaluated before the cutoff: rejected pairs neither consume K nor enter the cutoff-tie group, while their successors remain eligible for heap expansion. The search stops after completing the K-th admissible score-tie group or exhausting the Cartesian product.

Algorithm 2 MixOPD with standard OPD

Require: Dataset $\mathcal{Q}$; student π_θ; fixed teacher π_{tea}; sampled-log-ratio clip C_A; PPO clip ϵ_{clip}; policy epochs E_{PPO}; search parameters $V, M_{\mathrm{c}}, K, s_{\min} = 16$

for rollout minibatch $\mathbf{q} \sim \mathcal{Q}$ **do**
  $\pi_{\theta_{\mathrm{old}}} \leftarrow \mathrm{sg}(\pi_\theta)$; sample $\mathbf{x} \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid \mathbf{q})$
  Form $c_t = (\mathbf{q}, \mathbf{x}_{<t})$, token masks m_t, and pairwise-disjoint parsed spans $\{(b_j, a_j, \mathcal{I}(b_j), \{t_{j,h,r}\}, \{\ell_{j,h}\})\}_{j=1}^{J}$
  Construct $\widetilde{c}_t = (c_t, \xi_t^{\mathrm{V}}, \xi_t^{\mathrm{T}})$ for all t, and teacher-force once to cache $q_t = \pi_{\mathrm{tea}}(\cdot \mid \widetilde{c}_t)$
  $\mathbf{p}_{\mathbf{x}}^{\mathrm{old}} \leftarrow [\pi_{\theta_{\mathrm{old}}}(x_t \mid c_t)]_t$, $\mathbf{q}_{\mathbf{x}} \leftarrow [q_t(x_t)]_t$
  $\widehat{\mathbf{A}} \leftarrow \mathrm{sg}\big(\mathrm{clip}(\log \mathbf{q}_{\mathbf{x}} - \log \mathbf{p}_{\mathbf{x}}^{\mathrm{old}}, -C_A, C_A)\big)$; $\mathbf{w} \leftarrow \mathrm{ones_like}(\mathbf{p}_{\mathbf{x}}^{\mathrm{old}})$
  for $j = 1, \ldots, J$ **do**
    Form search record $\mathcal{R}_j$ from b_j, $\{t_{j,h,r}\}$, $\{\ell_{j,h}\}$, and the cached teacher laws
    $(\mathcal{B}_{\mathrm{tea},j}^{K}, \widehat{q}_j, \widehat{\omega}_{j,K}) \leftarrow \textsc{TeacherForcedCoordinateSearch}(\mathcal{R}_j, V, M_{\mathrm{c}}, K, s_{\min})$
    $\mathbf{w}[\mathcal{I}(b_j)] \leftarrow \mathrm{sg}(\widehat{\omega}_{j,K})$
  for $e = 1, \ldots, E_{\mathrm{PPO}}$ **do**
    for each policy-update minibatch $\mathcal{M}$ **do**
      $\boldsymbol{\varrho}_{\mathcal{M}} \leftarrow [\pi_\theta(x_t \mid c_t)]_{t \in \mathcal{M}} / \mathbf{p}_{\mathbf{x},\mathcal{M}}^{\mathrm{old}}$
      $\boldsymbol{\ell}_{\mathcal{M}}^{\mathrm{OPD}} \leftarrow \mathrm{maximum}\Big(-\boldsymbol{\varrho}_{\mathcal{M}} \odot \widehat{\mathbf{A}}_{\mathcal{M}}, -\mathrm{clip}(\boldsymbol{\varrho}_{\mathcal{M}}, 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}}) \odot \widehat{\mathbf{A}}_{\mathcal{M}}\Big)$
      $\mathcal{L} \leftarrow \mathrm{sum}(\mathbf{m}_{\mathcal{M}} \odot \mathbf{w}_{\mathcal{M}} \odot \boldsymbol{\ell}_{\mathcal{M}}^{\mathrm{OPD}}) / \mathrm{sum}(\mathbf{m}_{\mathcal{M}})$; update θ with $\nabla_\theta \mathcal{L}$

Algorithm 3 MixOPD with EOPD

Require: Dataset $\mathcal{Q}$; student π_θ; fixed teacher π_{tea}; sampled-log-ratio clip C_A; PPO clip ϵ_{clip}; policy epochs E_{PPO}; token support L; entropy threshold η_H; FKL weight λ_{F}; search parameters $V, M_{\mathrm{c}}, K, s_{\min} = 16$

for rollout minibatch $\mathbf{q} \sim \mathcal{Q}$ **do**
  $\pi_{\theta_{\mathrm{old}}} \leftarrow \mathrm{sg}(\pi_\theta)$; sample $\mathbf{x} \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid \mathbf{q})$
  Form $c_t = (\mathbf{q}, \mathbf{x}_{<t})$, token masks m_t, and parsed spans $\{(b_j, a_j, \mathcal{I}(b_j), \{t_{j,h,r}\}, \{\ell_{j,h}\})\}_{j=1}^{J}$
  Construct $\widetilde{c}_t = (c_t, \xi_t^{\mathrm{V}}, \xi_t^{\mathrm{T}})$ for all t, and teacher-force once to cache $q_t = \pi_{\mathrm{tea}}(\cdot \mid \widetilde{c}_t)$
  $\mathbf{p}_{\mathbf{x}}^{\mathrm{old}} \leftarrow [\pi_{\theta_{\mathrm{old}}}(x_t \mid c_t)]_t$, $\mathbf{q}_{\mathbf{x}} \leftarrow [q_t(x_t)]_t$
  $\widehat{\mathbf{A}} \leftarrow \mathrm{sg}\big(\mathrm{clip}(\log \mathbf{q}_{\mathbf{x}} - \log \mathbf{p}_{\mathbf{x}}^{\mathrm{old}}, -C_A, C_A)\big)$; $\mathbf{w} \leftarrow \mathrm{ones_like}(\mathbf{p}_{\mathbf{x}}^{\mathrm{old}})$
  for each valid response position t **do**
    $g_t \leftarrow \mathrm{sg}(\mathbf{1}\{H(q_t) > \eta_H\})$; $\mathcal{S}_t^L \leftarrow \mathrm{topk}(q_t, L).\mathrm{indices}$
    $\bar{q}_t \leftarrow \mathrm{sg}\big(q_t[\mathcal{S}_t^L] / \mathrm{sum}(q_t[\mathcal{S}_t^L])\big)$
  for $j = 1, \ldots, J$ **do**
    Form search record $\mathcal{R}_j$ from b_j, $\{t_{j,h,r}\}$, $\{\ell_{j,h}\}$, and the cached teacher laws
    $(\mathcal{B}_{\mathrm{tea},j}^{K}, \widehat{q}_j, \widehat{\omega}_{j,K}) \leftarrow \textsc{TeacherForcedCoordinateSearch}(\mathcal{R}_j, V, M_{\mathrm{c}}, K, s_{\min})$
    $\mathbf{w}[\mathcal{I}(b_j)] \leftarrow \mathrm{sg}(\widehat{\omega}_{j,K})$
  for $e = 1, \ldots, E_{\mathrm{PPO}}$ **do**
    for each policy-update minibatch $\mathcal{M}$ **do**
      $\boldsymbol{\varrho}_{\mathcal{M}} \leftarrow [\pi_\theta(x_t \mid c_t)]_{t \in \mathcal{M}} / \mathbf{p}_{\mathbf{x},\mathcal{M}}^{\mathrm{old}}$
      $\boldsymbol{\ell}_{\mathcal{M}}^{\mathrm{OPD}} \leftarrow \mathrm{maximum}\Big(-\boldsymbol{\varrho}_{\mathcal{M}} \odot \widehat{\mathbf{A}}_{\mathcal{M}}, -\mathrm{clip}(\boldsymbol{\varrho}_{\mathcal{M}}, 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}}) \odot \widehat{\mathbf{A}}_{\mathcal{M}}\Big)$
      $\mathbf{F}_{\mathcal{M}}(\theta) \leftarrow \Big[\sum_{v \in \mathcal{S}_t^L} \bar{q}_t(v) \log \frac{\bar{q}_t(v)}{\pi_\theta(v \mid c_t)}\Big]_{t \in \mathcal{M}}$
      $\mathbf{S}_{\mathcal{M}}^{(\lambda_{\mathrm{F}})} \leftarrow \boldsymbol{\ell}_{\mathcal{M}}^{\mathrm{OPD}} + \lambda_{\mathrm{F}} \mathbf{g}_{\mathcal{M}} \odot \mathbf{F}_{\mathcal{M}}(\theta)$
      $\mathcal{L} \leftarrow \mathrm{sum}(\mathbf{m}_{\mathcal{M}} \odot \mathbf{w}_{\mathcal{M}} \odot \mathbf{S}_{\mathcal{M}}^{(\lambda_{\mathrm{F}})}) / \mathrm{sum}(\mathbf{m}_{\mathcal{M}})$; update θ with $\nabla_\theta \mathcal{L}$

C Experimental Setup

C.1 Benchmarks and metrics

Our evaluation targets vision-centric reasoning from two complementary aspects: (i) *general-purpose* visual tasks that probe broad perception, knowledge, and vision-indispensable reasoning; and (ii) *fine-grained* visual understanding that stresses high-resolution inputs and small evidence, where explicit evidence localization is expected to be most beneficial. We report results on *seven* widely used benchmarks, following their default protocols. We use accuracy for answer correctness and additionally report localization quality (*i.e.*, mIoU) when annotated evidence is available.

General-purpose Benchmarks. MMStar Chen et al. (2024) is a *vision-indispensable* benchmark of 1,500 carefully curated samples, each filtered to genuinely require the image. It spans *six* core capabilities: coarse perception, fine-grained perception, instance reasoning, logical reasoning, science & technology, and mathematics, and is reported with *accuracy* as a metric. **CV-Bench** Tong et al. (2024) is a vision-centric VQA suite of 2,638 human-inspected examples repurposed from classic vision datasets. It probes fundamental *2D* understanding (*e.g.*, spatial relationships and counting) and *3D* understanding (*e.g.*, depth ordering and relative distance), measuring perception that demands real visual evidence rather than text shortcuts. **MME-RealWorld** Zhang et al. (2025b) is a large-scale, high-resolution, real-world benchmark covering *five* practical domains: OCR, remote sensing, diagram & table, monitoring, and autonomous driving. It provides both *Chinese* and *English* versions, *i.e.*, $\text{MME}^{\text{rw}}_{\text{CN}}$ and $\text{MME}^{\text{rw}}_{\text{EN}}$, where the Chinese version incorporates additional images focusing on Chinese-specific scenarios and is re-annotated by native volunteers, resulting in 5,917 QA pairs.

Fine-grained Benchmarks. V* Bench Wu & Xie (2024) is a fine-grained visual understanding benchmark built on high-resolution scenes (191 samples) that emphasizes small targets. It comprises an *Attribute* subset and a *Spatial* subset; we report accuracy overall and per subset. **ZoomBench** Wei et al. (2026) is a fine-grained perception benchmark, which includes 845 high-quality VQA samples over high-resolution images spanning *six* dimensions, *e.g.*, color attributes, fine-grained counting, structural attributes, OCR, material attributes, and object identification. **HR-Bench (4K/8K)** Wang et al. (2025c) evaluates high-resolution VQA under long-context visual inputs at 4K and 8K resolutions. It contains both a *Single*-instance setting (*e.g.*, attribute recognition and OCR) and a *Cross*-instance setting (*e.g.*, spatial relations), stressing the ability to preserve fine details across large visual fields; we report accuracy for each resolution and setting. **TreeBench** Wang et al. (2025a) is a grounded reasoning benchmark, jointly evaluating *answer correctness* and *evidence localization quality* (mIoU). Its taxonomy separates *Perception* (*e.g.*, attributes, material, physical state, OCR) from *Reasoning* (*e.g.*, perspective transform, ordering, contact & occlusion), enabling a targeted analysis of whether a method improves perception behaviors, reasoning behaviors, or both.

C.2 Baselines

We compare MixOPD against *five* types of baselines: (i) *agentic frameworks* with explicit tool-use/code-execution; (ii) leading *proprietary* LVLMs; (iii) *open-source* LVLMs across scales; (iv) *distillation-based alternatives*; and (v) *representative post-training paradigm* re-implemented under an identical backbone and data budget to isolate the effect of the optimization strategy.

Agentic Frameworks enhance LVLMs with multi-turn interaction loops with sand-boxed tool use or external code executors. **DeepEyes** Zheng et al. (2025) incentivizes *thinking with images* via end-to-end RL, learning to interleave reasoning with tools for image manipulations, *e.g.*, zoom-in. Its successor **DeepEyesV2** Hong et al. (2025a) moves toward a fully agentic multimodal model that invokes code execution and web search. **Thyme** Zhang et al. (2025a) realizes *think beyond images* by autonomously generating and executing code in a sandbox to manipulate images, *e.g.*, crop, zoom, enhance. **PFlowNet** Li et al. (2026a) decouples perception from reasoning through a self-conditioned generation framework and achieves efficiently visually grounded reasoning via a structured *perceptual flow* instead of explicit tool calls. **SenseNova MARS** Chng et al. (2025) is an agentic multimodal reasoning system that dynamically composes image-crop and search tools to strengthen evidence acquisition and multi-step grounded reasoning.

Private LVLMs. We report the results from frontier private LVLMs queried through their official APIs under the corresponding benchmark protocols, providing an upper-bound reference for general and fine-grained visual understanding. **GPT-5.2** OpenAI (2025) and **GPT-5.4** OpenAI (2026) are

frontier multimodal models in OpenAI's GPT-5 lineage, designed for impressive multimodal agentic intelligence. **Gemini-3.1 Pro** DeepMind (2026a) and **Gemini-3.5 Flash** DeepMind (2026b) are natively multimodal, long-context models from Google DeepMind's Gemini 3 lineage, where the *Pro* tier targets high-capability reasoning and the *Flash* tier targets efficiency and throughput.

Open-Source LVLMs. We also consider influential LVLMs widely recognized by the community across multiple scales. **GLM-4.6V** Hong et al. (2025b) is an open mixture-of-experts (MoE) vision-language model in the GLM-4.5V/4.6 lineage, featuring an explicit *thinking* mode and native multimodal tool-calling. The **Qwen3-VL series** Bai et al. (2025a) is a strong open multimodal family that incorporates architectural improvements such as *DeepStack* and *interleaved multimodal position encoding* for fine-grained perception and long-context grounding. **Kimi-K2.6** Kimi Team (2026) is a large native-multimodal MoE model in Moonshot AI's Kimi-K2 lineage, positioned for long-horizon agentic and vision-centric tasks. **MiMo-VL RL** Yue et al. (2025) is a compact ($\sim$7B) reasoning-oriented VLM trained with mixed on-policy RL, which is competitive with substantially larger models on perception- and reasoning-heavy tasks. **MiniCPM-V 4.5** Yu et al. (2026b) is an efficient on-device-friendly MLLM that combines high-resolution image slicing with a hybrid fast/deep thinking mode, delivering strong perception and OCR performance per parameter.

Distillation-based Alternatives aim to internalize fine-grained visual manipulation within a single forward pass. **ZwZ** Wei et al. (2026) moves zooming from an inference-time tool call to a *region-to-image* distillation objective, enabling fine-grained perception on full images in a single forward pass. **VisionOPD** Yuan et al. (2026) performs regional-to-global on-policy self-distillation, where the same model acts as a teacher conditioned on the zoomed-in evidence view and as a student conditioned on the full image, transferring fine-detail perception via token-level distillation without external teachers or labels. **ImagineOPD** Cai et al. (2026) internalizes visual manipulation through on-policy self-distillation: a teacher that *thinks with images* guides a student to *imagine* the relevant visual evidence and reason about it internally, removing the need for external image operations at inference.

Representative RL Methods. We finally compare our method against representative post-training algorithms re-implemented on the same backbone and training data. **GRPO** Guo et al. (2025) is a critic-free policy-optimization algorithm that estimates advantages by group normalization over multiple sampled responses, removing the value network. **DAPO** Yu et al. (2026a) scales and stabilizes GRPO-style training through decoupled (clip-higher) clipping, dynamic sampling of informative prompts, a token-level policy-gradient loss, and overlong-response reward shaping. **OPD** Lu & Lab (2025) combines the on-policy data distribution of RL with the dense, per-token supervision of distillation: the student samples its own rollouts while a teacher scores them token by token, minimizing a per-token reverse KL that supervises every step rather than only the final reward. **EOPD** Jin et al. (2026) augments OPD's reverse-KL objective with an additive forward-KL term at positions where the teacher's token-level entropy exceeds a threshold, combining mode-seeking supervision throughout with mode-covering guidance at uncertain steps to preserve the teacher's distributional diversity. All methods are trained on the same dataset described in Appendix B.1, and the detailed training recipes for the baselines are provided in Appendix B.2.

C.3 Evaluation Protocol

For a controlled comparison, we evaluate every open-source baseline from its official checkpoint through a single inference pipeline built on vLLM 0.8.0 and $8\times$ NVIDIA H800 GPUs, removing confounds from heterogeneous inference stacks. VisionOPD is natively built on Qwen3.5 4B/9B; to disentangle method from backbone, we reproduce it from its official codebase and recipe but instantiate it on Qwen3-VL 4B/8B, reporting these backbone-matched variants in Table 1. Proprietary systems (*e.g.*, GPT-5, Gemini-3 series) are queried through their official APIs. Unless noted otherwise, all models share one decoding configuration with temperature 0.7, top-p 0.8, top-k 20, presence penalty 1.5, repetition penalty 1.0, capped at 8192 generated tokens—following each baseline's official system prompt and template where available. We score with a two-stage protocol: predictions are first matched to the ground truth by symbolic equivalence (via `MathRuler`) plus option-letter matching on multiple-choice benchmarks, and only the residual cases are deferred to an LLM judge (Qwen3-30B-A3B-Instruct-2507, greedy decoding) for semantic equivalence—keeping scoring largely rule-based and reproducible yet robust to surface form. All results are averaged over three independent runs with seeds $\{41, 3047, 2027\}$. Unless stated otherwise, MixOPD denotes its standard-OPD instantiation in §4.1) and ablation study adopts its 8B variant.

D Additional Empirical Results

D.1 Additional Ablation Studies

Table 8: Ablation study on teacher parameter size under OPD (*left*) and EOPD (*right*) formulations.

	mIoU	TreeB	VStar	Zoom	$\text{MME}^{\text{rw}}_{\text{EN}}$
8B	0.44	50.70	89.88	**58.11**	67.82
32B	**0.46**	**53.25**	**90.40**	57.83	68.27
235B	0.39	49.06	87.61	55.66	**68.74**

	mIoU	TreeB	VStar	Zoom	$\text{MME}^{\text{rw}}_{\text{EN}}$
8B	0.45	52.10	89.70	57.00	68.90
32B	**0.48**	**53.99**	**92.50**	**59.76**	69.98
235B	0.44	51.36	89.18	59.13	**70.16**

Teacher Scale. As shown in Table 8, we observe no clear monotonic gain from scaling the teacher. This differs from STEM-style reasoning tasks, where larger teachers often provide stronger learning signals via more complex reasoning traces. Actually, most visual reasoning tasks are bottlenecked by grounding: once the task-relevant evidence is localized, the remaining reasoning is straightforward. This weakens the advantage of larger teachers in distillation. Moreover, the teacher-side privileged inputs further narrow the capability gap across teacher scales.

Table 9: Ablation study on the strategy for handling invalid grounding spans of students.

	mIoU	TreeB	VStar	Zoom	$\text{HR}_{4\text{K}}$	$\text{HR}_{8\text{K}}$	$\text{MME}^{\text{rw}}_{\text{CN}}$	$\text{MME}^{\text{rw}}_{\text{EN}}$
Under OPD formulation								
Sample-level Skip	0.43	51.85	89.88	57.55	83.04	80.46	66.18	67.66
Span-level Degrade	**0.46**$_{+0.03}$	**53.25**$_{+1.40}$	**90.40**$_{+0.52}$	**57.83**$_{+0.28}$	**84.50**$_{+1.46}$	**80.92**$_{+0.46}$	**66.53**$_{+0.35}$	**68.27**$_{+0.61}$
Under EOPD formulation								
Sample-level Skip	0.45	52.02	90.92	58.15	84.33	**81.50**	67.60	69.54
Span-level Degrade	**0.48**$_{+0.03}$	**53.99**$_{+1.97}$	**92.50**$_{+1.58}$	**59.76**$_{+1.61}$	**85.25**$_{+0.92}$	81.13$_{-0.37}$	**68.10**$_{+0.50}$	**69.98**$_{+0.44}$

Invalid Span Handling. When a student grounding span cannot be parsed into a valid bounding box, *e.g.*, due to missing digits, malformed special tokens, out-of-range coordinates, or invalid coordinate ordering, we consider two handling strategies: (1) *Sample-level Skip*, which excludes the entire sample from loss computation; and (2) *Span-level Degrade*, which locally replaces the spatially calibrated objective for the invalid span with the standard objective. We compare these strategies in Table 9. *Span-level Degrade* almost performs better, which we attribute to its higher sample efficiency. In samples involving multiple pieces of visual evidence, only one or two grounding spans may be invalid; discarding the entire sample would also remove valid supervision from all remaining token positions. Empirically, *Sample-level Skip* excludes approximately 7% of the training samples.

Table 10: Controlled ablation of unreliable token-level supervision localization and calibration.

	mIoU	TreeB	VStar	Zoom	$\text{HR}_{4\text{K}}$	$\text{HR}_{8\text{K}}$	$\text{MME}^{\text{rw}}_{\text{CN}}$	$\text{MME}^{\text{rw}}_{\text{EN}}$
Fixing the recalibration objective: Forward KL								
Special Tags	**0.43**	50.12	89.35	54.20	82.46	**81.38**	64.32	63.74
Entropy Gate (EOPD)	0.42$_{-0.01}$	**51.52**$_{+1.40}$	**89.70**$_{+0.35}$	**54.36**$_{+0.16}$	**83.25**$_{+0.79}$	80.96$_{-0.42}$	**66.46**$_{+2.14}$	**67.87**$_{+4.13}$
Fixing the supervision localization strategy: Special Tags								
Forward KL	0.43	50.12	89.35	54.20	82.46	**81.38**	64.32	63.74
Spatial Div (MixOPD)	**0.46**$_{+0.03}$	**53.25**$_{+3.13}$	**90.40**$_{+1.05}$	**57.83**$_{+3.63}$	**84.50**$_{+2.04}$	80.92$_{-0.46}$	**66.53**$_{+2.21}$	**68.27**$_{+4.53}$

Methodological Design. EOPD and MixOPD both recalibrate the token-level reverse-KL supervision in OPD, but differ along two design axes: *where* unreliable supervision is localized and *how* it is recalibrated. Table 10 disentangles these choices. With forward KL fixed, entropy gating improves six of the seven downstream metrics over special-tag localization, with pronounced gains on $\text{MME}^{\text{rw}}_{\text{EN}}$ ($+4.13$), $\text{MME}^{\text{rw}}_{\text{CN}}$ ($+2.14pp.$), and TreeB ($+1.40pp.$), while leaving mIoU essentially unchanged. This pattern is consistent with the analysis in Sec. 4.4: high-entropy positions include not only grounding tokens but also uncertain reasoning decisions that can redirect subsequent trajectories and are not exposed by grounding tags. Conversely, when localization is fixed to tagged grounding spans, spatial divergence improves seven of the eight reported metrics over forward KL, raising mIoU from 0.43 to 0.46, TreeB by $3.13pp.$ and Zoom by $3.63pp.$. Forward KL broadens the coverage of teacher-supported alternatives at each autoregressive position, but provides no explicit criterion for whether the resulting coordinate sequence constitutes a geometrically valid region. Spatial divergence instead evaluates the realized proposal as a whole in image space, providing span-level credit aligned with grounding quality. These comparisons demonstrate that EOPD and MixOPD address complementary failure modes of OPD in vision-centric reasoning.

D.2 SIGNIFICANCE ANALYSIS

Table 11: Raw paired correctness counts for MixOPD and competitive alternatives across random seeds and benchmarks. Results are reported for 4B and 8B models, where A denotes MixOPD, B denotes the compared method, and subscripts c and w indicate correct and incorrect predictions, respectively. The discordant pairs (A_c, B_w) and (A_w, B_c) are used for the two-sided McNemar test.

Seed	VStar	Zoom	HR4K	HR8K	MME-CN	MME-EN	MMStar	CV-B
MixOPD 4B (A) vs. Qwen3-VL 4B (B)								
Seed=41	B_c B_w A_c 146 25 A_w 8 12	B_c B_w A_c 317 148 A_w 17 363	B_c B_w A_c 587 60 A_w 31 122	B_c B_w A_c 559 76 A_w 30 135	B_c B_w A_c 3607 342 A_w 190 1778	B_c B_w A_c 14433 1640 A_w 760 6766	B_c B_w A_c 997 53 A_w 60 390	B_c B_w A_c 2113 184 A_w 112 229
Seed=3047	B_c B_w A_c 146 25 A_w 8 12	B_c B_w A_c 330 133 A_w 18 364	B_c B_w A_c 598 61 A_w 32 109	B_c B_w A_c 549 85 A_w 29 137	B_c B_w A_c 3515 338 A_w 186 1878	B_c B_w A_c 14127 1978 A_w 744 6750	B_c B_w A_c 996 53 A_w 62 389	B_c B_w A_c 2142 185 A_w 113 198
Seed=2027	B_c B_w A_c 143 29 A_w 8 11	B_c B_w A_c 327 149 A_w 18 351	B_c B_w A_c 588 60 A_w 31 121	B_c B_w A_c 547 89 A_w 29 135	B_c B_w A_c 3604 342 A_w 190 1781	B_c B_w A_c 14445 1995 A_w 761 6398	B_c B_w A_c 1002 68 A_w 53 377	B_c B_w A_c 2147 186 A_w 114 191
MixOPD 4B (A) vs. ZwZ 4B (B)								
Seed=41	B_c B_w A_c 159 12 A_w 9 11	B_c B_w A_c 431 34 A_w 23 357	B_c B_w A_c 608 39 A_w 32 121	B_c B_w A_c 603 32 A_w 37 128	B_c B_w A_c 3630 319 A_w 192 1776	B_c B_w A_c 14893 1180 A_w 784 6742	B_c B_w A_c 997 53 A_w 60 390	B_c B_w A_c 2182 115 A_w 127 214
Seed=3047	B_c B_w A_c 159 12 A_w 9 11	B_c B_w A_c 429 34 A_w 23 359	B_c B_w A_c 611 48 A_w 33 108	B_c B_w A_c 602 32 A_w 37 129	B_c B_w A_c 3660 193 A_w 258 1806	B_c B_w A_c 14587 1518 A_w 768 6726	B_c B_w A_c 996 53 A_w 62 389	B_c B_w A_c 2210 117 A_w 129 182
Seed=2027	B_c B_w A_c 158 14 A_w 9 10	B_c B_w A_c 441 35 A_w 24 345	B_c B_w A_c 615 33 A_w 40 112	B_c B_w A_c 604 32 A_w 37 127	B_c B_w A_c 3705 241 A_w 196 1775	B_c B_w A_c 14905 1535 A_w 785 6374	B_c B_w A_c 1002 68 A_w 53 377	B_c B_w A_c 2189 144 A_w 116 189
MixOPD 4B (A) vs. VisionOPD 4B (B)								
Seed=41	B_c B_w A_c 150 21 A_w 8 12	B_c B_w A_c 341 124 A_w 18 362	B_c B_w A_c 601 46 A_w 32 121	B_c B_w A_c 560 75 A_w 30 135	B_c B_w A_c 3711 238 A_w 196 1772	B_c B_w A_c 15269 804 A_w 1106 6420	B_c B_w A_c 994 56 A_w 53 397	B_c B_w A_c 2161 136 A_w 114 227
Seed=3047	B_c B_w A_c 152 19 A_w 8 12	B_c B_w A_c 354 109 A_w 19 363	B_c B_w A_c 589 70 A_w 32 109	B_c B_w A_c 570 64 A_w 31 135	B_c B_w A_c 3660 193 A_w 233 1831	B_c B_w A_c 15250 855 A_w 803 6691	B_c B_w A_c 993 56 A_w 53 398	B_c B_w A_c 2159 168 A_w 114 197
Seed=2027	B_c B_w A_c 149 23 A_w 8 11	B_c B_w A_c 342 134 A_w 18 351	B_c B_w A_c 587 61 A_w 31 121	B_c B_w A_c 572 64 A_w 31 133	B_c B_w A_c 3714 232 A_w 196 1775	B_c B_w A_c 15568 872 A_w 820 6339	B_c B_w A_c 996 74 A_w 53 377	B_c B_w A_c 2127 206 A_w 112 193
MixOPD 4B (A) vs. ImagineOPD 4B (B)								
Seed=41	B_c B_w A_c 138 33 A_w 20 0	B_c B_w A_c 369 96 A_w 9 371	B_c B_w A_c 579 68 A_w 55 98	B_c B_w A_c 609 26 A_w 31 134	B_c B_w A_c 3479 470 A_w 362 1606	B_c B_w A_c 13820 2253 A_w 1807 5719	B_c B_w A_c 888 162 A_w 98 352	B_c B_w A_c 1944 353 A_w 330 11
Seed=3047	B_c B_w A_c 138 33 A_w 20 0	B_c B_w A_c 371 92 A_w 8 374	B_c B_w A_c 595 64 A_w 40 101	B_c B_w A_c 621 13 A_w 20 146	B_c B_w A_c 3514 339 A_w 307 1757	B_c B_w A_c 14329 1776 A_w 1300 6194	B_c B_w A_c 867 182 A_w 140 311	B_c B_w A_c 2047 280 A_w 230 81
Seed=2027	B_c B_w A_c 139 33 A_w 19 0	B_c B_w A_c 357 119 A_w 11 358	B_c B_w A_c 573 75 A_w 61 91	B_c B_w A_c 601 35 A_w 26 138	B_c B_w A_c 3451 495 A_w 381 1590	B_c B_w A_c 14259 2181 A_w 1667 5492	B_c B_w A_c 853 217 A_w 128 302	B_c B_w A_c 2061 272 A_w 211 94
MixOPD 8B (A) vs. Qwen3-VL 8B (B)								
Seed=41	B_c B_w A_c 152 21 A_w 9 9	B_c B_w A_c 350 147 A_w 19 329	B_c B_w A_c 612 73 A_w 33 82	B_c B_w A_c 577 67 A_w 31 125	B_c B_w A_c 3591 349 A_w 189 1788	B_c B_w A_c 14316 1566 A_w 754 6963	B_c B_w A_c 1021 56 A_w 54 369	B_c B_w A_c 2126 192 A_w 112 208
Seed=3047	B_c B_w A_c 154 19 A_w 9 9	B_c B_w A_c 352 131 A_w 19 343	B_c B_w A_c 608 65 A_w 33 94	B_c B_w A_c 576 79 A_w 31 114	B_c B_w A_c 3604 322 A_w 190 1801	B_c B_w A_c 14622 1582 A_w 770 6625	B_c B_w A_c 999 101 A_w 53 347	B_c B_w A_c 2154 162 A_w 114 208
Seed=2027	B_c B_w A_c 152 20 A_w 8 11	B_c B_w A_c 340 146 A_w 18 341	B_c B_w A_c 611 59 A_w 33 97	B_c B_w A_c 565 78 A_w 30 127	B_c B_w A_c 3515 428 A_w 186 1788	B_c B_w A_c 14665 1584 A_w 772 6578	B_c B_w A_c 1019 60 A_w 54 367	B_c B_w A_c 2122 160 A_w 112 244
MixOPD 8B (A) vs. ZwZ 8B (B)								
Seed=41	B_c B_w A_c 158 15 A_w 9 9	B_c B_w A_c 449 48 A_w 24 324	B_c B_w A_c 643 42 A_w 34 81	B_c B_w A_c 611 33 A_w 49 107	B_c B_w A_c 3743 197 A_w 237 1740	B_c B_w A_c 15087 795 A_w 1159 6558	B_c B_w A_c 1023 54 A_w 63 360	B_c B_w A_c 2202 116 A_w 116 204
Seed=3047	B_c B_w A_c 158 15 A_w 9 9	B_c B_w A_c 458 25 A_w 29 333	B_c B_w A_c 639 34 A_w 34 93	B_c B_w A_c 622 33 A_w 37 108	B_c B_w A_c 3689 237 A_w 195 1796	B_c B_w A_c 15393 811 A_w 821 6574	B_c B_w A_c 1009 91 A_w 54 346	B_c B_w A_c 2200 116 A_w 116 206
Seed=2027	B_c B_w A_c 157 15 A_w 9 10	B_c B_w A_c 459 27 A_w 25 334	B_c B_w A_c 628 42 A_w 34 96	B_c B_w A_c 610 33 A_w 51 106	B_c B_w A_c 3705 238 A_w 196 1778	B_c B_w A_c 15436 813 A_w 823 6527	B_c B_w A_c 1025 54 A_w 59 362	B_c B_w A_c 2167 115 A_w 155 201
MixOPD 8B (A) vs. VisionOPD 8B (B)								
Seed=41	B_c B_w A_c 153 20 A_w 9 9	B_c B_w A_c 377 120 A_w 20 328	B_c B_w A_c 608 77 A_w 32 83	B_c B_w A_c 572 72 A_w 31 125	B_c B_w A_c 3663 277 A_w 193 1784	B_c B_w A_c 14520 1362 A_w 765 6952	B_c B_w A_c 1018 59 A_w 54 369	B_c B_w A_c 2137 181 A_w 113 207
Seed=3047	B_c B_w A_c 153 20 A_w 9 9	B_c B_w A_c 363 120 A_w 20 342	B_c B_w A_c 604 69 A_w 32 95	B_c B_w A_c 573 82 A_w 31 114	B_c B_w A_c 3572 354 A_w 188 1803	B_c B_w A_c 14826 1378 A_w 781 6614	B_c B_w A_c 1017 83 A_w 54 346	B_c B_w A_c 2139 177 A_w 113 209
Seed=2027	B_c B_w A_c 152 20 A_w 9 10	B_c B_w A_c 376 110 A_w 20 339	B_c B_w A_c 593 77 A_w 32 98	B_c B_w A_c 585 58 A_w 31 126	B_c B_w A_c 3588 355 A_w 189 1785	B_c B_w A_c 14869 1380 A_w 783 6567	B_c B_w A_c 1020 59 A_w 54 367	B_c B_w A_c 2133 149 A_w 113 243
MixOPD 8B (A) vs. ImagineOPD 8B (B)								
Seed=41	B_c B_w A_c 151 22 A_w 9 9	B_c B_w A_c 344 153 A_w 56 292	B_c B_w A_c 576 109 A_w 41 74	B_c B_w A_c 567 77 A_w 56 100	B_c B_w A_c 3529 411 A_w 340 1637	B_c B_w A_c 15261 621 A_w 495 7222	B_c B_w A_c 880 197 A_w 104 319	B_c B_w A_c 2168 150 A_w 122 198
Seed=3047	B_c B_w A_c 149 24 A_w 7 11	B_c B_w A_c 337 146 A_w 62 300	B_c B_w A_c 606 67 A_w 20 107	B_c B_w A_c 559 96 A_w 51 94	B_c B_w A_c 3576 350 A_w 313 1678	B_c B_w A_c 15695 509 A_w 367 7028	B_c B_w A_c 854 246 A_w 127 273	B_c B_w A_c 2075 241 A_w 218 104
Seed=2027	B_c B_w A_c 153 19 A_w 7 12	B_c B_w A_c 340 146 A_w 58 301	B_c B_w A_c 574 96 A_w 53 77	B_c B_w A_c 542 101 A_w 68 89	B_c B_w A_c 3590 353 A_w 288 1686	B_c B_w A_c 15303 946 A_w 752 6598	B_c B_w A_c 867 212 A_w 140 281	B_c B_w A_c 2081 201 A_w 165 191

Table 12: Raw paired correctness counts comparing MixOPD against different OPD formulations across random seeds and benchmarks. Results are reported for standard OPD and EOPD built on Qwen3-VL-4B/8B, where A denotes MixOPD, B denotes the corresponding OPD formulation, and subscripts c and w indicate correct and incorrect predictions, respectively. The discordant pairs (A_c, B_w) and (A_w, B_c) are used for the exact two-sided McNemar test.

Seed	TreeB	VStar	Zoom	HR4K	HR8K	MME-CN	MME-EN	MMStar	CV-B
Our method (A) vs. Standard OPD (B) built on Qwen3-VL 4B									
Seed=41	A_c 178 25 A_w 10 192	A_c 152 19 A_w 8 12	A_c 391 74 A_w 21 359	A_c 614 33 A_w 35 118	A_c 570 65 A_w 31 134	A_c 3581 368 A_w 189 1779	A_c 15107 966 A_w 796 6730	A_c 1011 39 A_w 54 396	A_c 2143 154 A_w 113 228
Seed=3047	A_c 178 30 A_w 10 187	A_c 154 17 A_w 9 11	A_c 389 74 A_w 21 361	A_c 616 43 A_w 33 108	A_c 570 64 A_w 31 135	A_c 3555 298 A_w 188 1876	A_c 15300 805 A_w 662 6832	A_c 995 54 A_w 53 398	A_c 2149 178 A_w 114 197
Seed=2027	A_c 185 24 A_w 10 186	A_c 152 20 A_w 8 11	A_c 403 73 A_w 22 347	A_c 615 33 A_w 33 119	A_c 583 53 A_w 31 133	A_c 3662 284 A_w 193 1778	A_c 15097 1343 A_w 795 6364	A_c 994 76 A_w 53 377	A_c 2142 191 A_w 113 192
Our method (A) vs. EOPD (B) built on Qwen3-VL 4B									
Seed=41	A_c 193 18 A_w 11 183	A_c 161 11 A_w 9 10	A_c 420 53 A_w 23 349	A_c 627 40 A_w 33 100	A_c 589 37 A_w 32 142	A_c 3631 391 A_w 192 1703	A_c 15108 1067 A_w 796 6628	A_c 1012 62 A_w 54 372	A_c 2228 143 A_w 118 149
Seed=3047	A_c 193 18 A_w 11 183	A_c 162 13 A_w 9 7	A_c 434 53 A_w 23 335	A_c 613 54 A_w 33 100	A_c 594 32 A_w 41 133	A_c 3714 335 A_w 196 1672	A_c 15052 1064 A_w 793 6690	A_c 996 77 A_w 53 374	A_c 2189 141 A_w 116 192
Seed=2027	A_c 187 25 A_w 10 183	A_c 161 13 A_w 9 8	A_c 433 39 A_w 23 350	A_c 621 33 A_w 38 108	A_c 590 49 A_w 32 129	A_c 3608 329 A_w 190 1790	A_c 15366 1080 A_w 809 6344	A_c 995 79 A_w 53 373	A_c 2229 143 A_w 118 148
Our method (A) vs. Standard OPD (B) built on Qwen3-VL 8B									
Seed=41	A_c 190 26 A_w 10 179	A_c 156 17 A_w 9 9	A_c 421 76 A_w 23 325	A_c 632 53 A_w 34 81	A_c 608 36 A_w 32 124	A_c 3639 301 A_w 192 1785	A_c 15087 795 A_w 1002 6715	A_c 1020 57 A_w 54 369	A_c 2141 177 A_w 113 207
Seed=3047	A_c 190 25 A_w 10 180	A_c 159 14 A_w 9 9	A_c 423 60 A_w 23 339	A_c 632 41 A_w 34 93	A_c 608 47 A_w 32 113	A_c 3722 204 A_w 196 1795	A_c 14960 1244 A_w 788 6607	A_c 1036 64 A_w 55 345	A_c 2147 169 A_w 114 208
Seed=2027	A_c 196 20 A_w 11 178	A_c 158 14 A_w 9 10	A_c 422 64 A_w 23 336	A_c 620 50 A_w 33 97	A_c 607 36 A_w 32 125	A_c 3615 328 A_w 191 1783	A_c 15027 1222 A_w 791 6559	A_c 1020 59 A_w 54 367	A_c 2167 115 A_w 136 220
Our method (A) vs. EOPD (B) built on Qwen3-VL 8B									
Seed=41	A_c 196 25 A_w 11 173	A_c 162 15 A_w 9 5	A_c 439 65 A_w 24 317	A_c 624 53 A_w 33 90	A_c 611 33 A_w 37 119	A_c 3752 297 A_w 198 1670	A_c 15409 988 A_w 812 6390	A_c 1018 88 A_w 54 340	A_c 2158 143 A_w 114 223
Seed=3047	A_c 202 19 A_w 11 173	A_c 163 15 A_w 9 4	A_c 441 65 A_w 24 315	A_c 637 54 A_w 34 75	A_c 615 43 A_w 33 109	A_c 3726 350 A_w 197 1644	A_c 15086 1652 A_w 794 6067	A_c 1034 71 A_w 55 340	A_c 2165 177 A_w 114 182
Seed=2027	A_c 195 19 A_w 11 180	A_c 162 13 A_w 9 7	A_c 427 78 A_w 23 317	A_c 636 42 A_w 34 88	A_c 612 33 A_w 35 120	A_c 3728 236 A_w 197 1756	A_c 15152 1256 A_w 798 6393	A_c 1035 71 A_w 55 339	A_c 2157 193 A_w 114 174

Aggregate accuracy suppresses both the paired structure of evaluation and the variability induced by training seeds. We therefore ask two distinct questions: within each seed, are discordant outcomes asymmetric in favor of MixOPD; and across seeds, does this advantage replicate? We answer the first with an exact paired test and report the second as replication, without pooling dependent evaluations.

Exact Paired Inference. Consider one fixed model-scale–benchmark–seed comparison. Let $A_i, B_i \in \{0, 1\}$ denote the correctness indicators of MixOPD and the compared method on the same example i, respectively. Their paired outcomes form the 2×2 table

$$\begin{array}{c|cc} & B_c & B_w \\ \hline A_c & n_{11} & n_{10} \\ A_w & n_{01} & n_{00} \end{array}, \qquad n_{ab} = \sum_{i=1}^{N} \mathbb{1}[A_i = a, B_i = b].$$

Here, n_{11} and n_{00} are concordant pairs. Only the discordant cells enter the conditional McNemar statistic: n_{10} counts examples solved uniquely by MixOPD, whereas n_{01} counts those solved uniquely by the comparator. The accompanying tables report all four cells for every seed, model scale, comparator, and benchmark: the first compares MixOPD with competitive alternatives, while the second isolates the effect of spatial calibration under standard OPD and EOPD. These counts are sufficient to reproduce both the effect size and the exact paired test directly, without reconstructing either from rounded aggregate scores. Because McNemar's test requires a binary per-example outcome, we analyze TreeB accuracy but exclude TreeB mIoU.

McNemar's null hypothesis is marginal homogeneity,

$$H_0 : \Pr(A_i = 1, B_i = 0) = \Pr(A_i = 0, B_i = 1).$$

Conditional on the number of discordant pairs $m = n_{10} + n_{01}$, the two discordant directions are equiprobable:

$$n_{10} \mid m \sim \text{Binomial}\left(m, \frac{1}{2}\right).$$

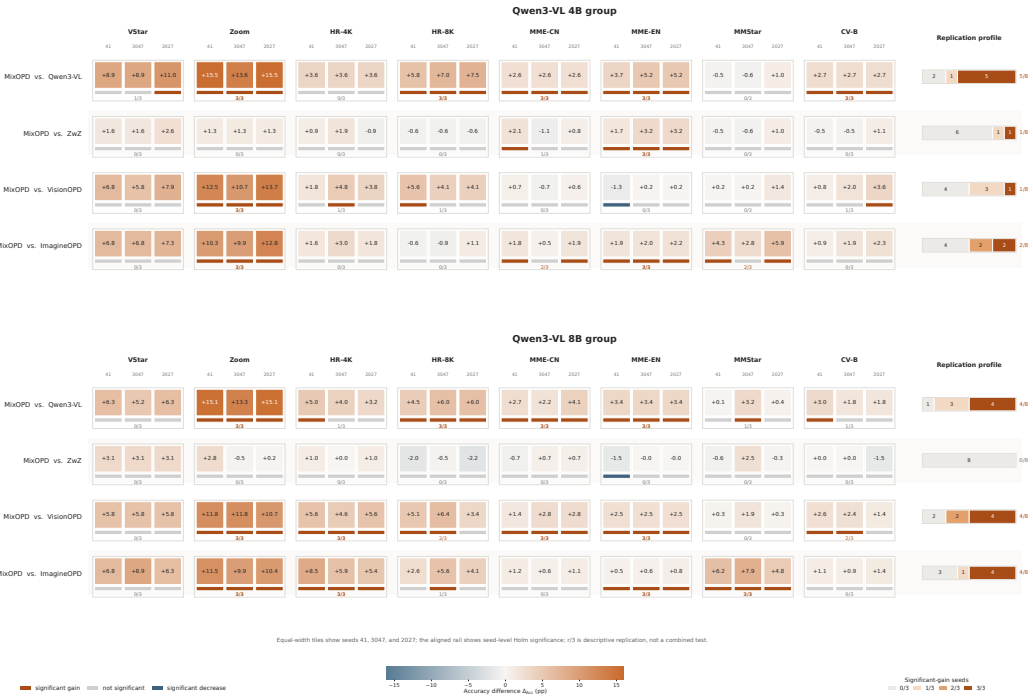

Figure 14: Significance of performance gain between MixOPD and competitive alternatives.

We compute the exact two-sided McNemar p-value as

$$p_{\mathrm{McN}} = \min\left\{1,\ 2\sum_{k=0}^{\min(n_{10},n_{01})}\binom{n_{10}+n_{01}}{k}2^{-(n_{10}+n_{01})}\right\}.$$

When $m = 0$, we set $p_{\mathrm{McN}} = 1$. This conditional calculation uses no large-sample χ^2 approximation or continuity correction, which is important for smaller benchmarks and comparisons with few discordant observations. The concordant cells do not enter the conditional statistic: n_{11} contributes to both accuracies, while n_{00} completes the table total. The signed effect follows from the same paired counts,

$$\Delta_{\mathrm{Acc}} = 100\,\frac{(n_{11}+n_{10})-(n_{11}+n_{01})}{N} = 100\,\frac{n_{10}-n_{01}}{N} \quad \text{percentage points.}$$

Multiplicity Control. We apply Holm correction separately for each prespecified analysis and model scale, jointly over all seed–comparator–benchmark hypotheses in that panel. Thus, at each model scale, the competitive-alternative family contains $3 \times 4 \times 8 = 96$ tests and $3 \times 2 \times 9 = 54$ tests for the formulation-controlled analysis. This controls the family-wise error rate for the broader panel-wise family, rather than for three seed-specific families, and remains valid under arbitrary dependence. We declare a panel-wise Holm-significant improvement only when

$$p^{\mathrm{adj}}_{\mathrm{Holm}} < 0.05 \qquad \text{and} \qquad \Delta_{\mathrm{Acc}} > 0.$$

A Holm-significant result with $\Delta_{\mathrm{Acc}} < 0$ is instead recorded as a significant decrease.

Seed-Resolved Visualization. Figures 14 and 15 visualize the competitive-alternative and formulation-controlled analyses, respectively. Each figure vertically aligns the 4B and 8B panels; rows denote the comparator or OPD formulation, and columns denote benchmarks. Every benchmark cell is partitioned into three equal-width tiles in the fixed seed order 41, 3047, and 2027. Tile fill encodes the signed Δ_{Acc} on a shared symmetric blue–white–orange scale within each figure, and the centered annotation reports the signed difference rounded to one decimal point. A thin rail beneath each tile separates inferential evidence from effect magnitude: orange denotes a Holm-significant gain, blue a Holm-significant decrease, and gray a non-significant difference.

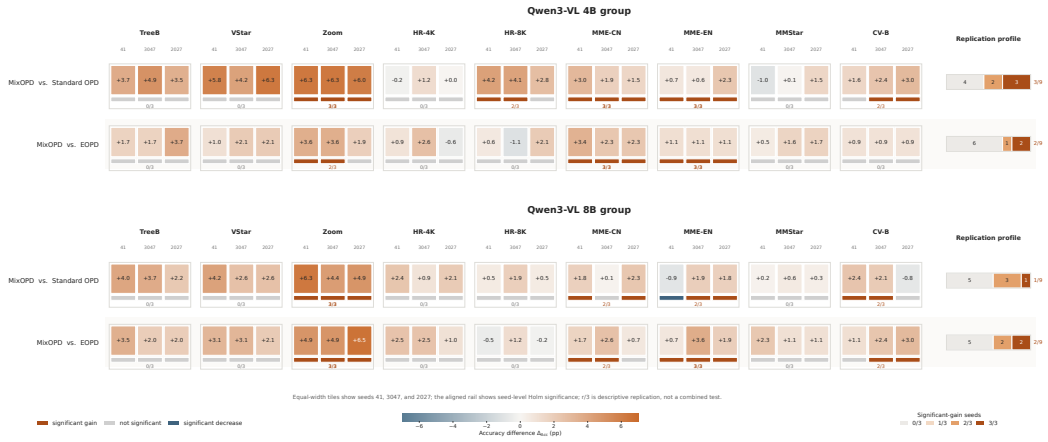

Figure 15: Significance of performance gains from spatial calibration under standard OPD and EOPD.

For each comparator–benchmark pair, we additionally report

$$r = \sum_{s \in \{41,3047,2027\}} \mathbb{1}\left[p^{\mathrm{adj}}_{s,\mathrm{Holm}} < 0.05 \ \wedge \ \Delta_{\mathrm{Acc},s} > 0\right],$$

shown as $r/3$ below the three tiles. Thus, r is specifically the Holm-significant-gain replication count; $r = 0$ does not imply inconsistent effect directions. The right marginal profile summarizes, for each row, how many benchmarks attain $r = 0, 1, 2$, or 3, followed by the number attaining full $3/3$ replication. This cross-seed layer is deliberately descriptive. Because the same benchmark examples are reused across seeds, their statistics are dependent; naively pooling the 2×2 tables or applying an independence-based p-value combination would not be justified without modeling this dependence.

Comparisons with Competitive Alternatives. Figure 14 reveals a structured, rather than uniform, advantage. Across the 192 seed-level comparisons, 163 favor MixOPD and 80 remain significant after Holm correction, compared with only two significant decreases; 21 of the 64 comparator–benchmark–scale cells replicate in all three seeds. Zoom provides the clearest case, reaching $3/3$ replication against Qwen3-VL, VisionOPD, and ImagineOPD at both model scales, with mean gains of approximately 11–15 points. Against ImagineOPD specifically, 46 of the 48 seed-level comparisons favor MixOPD and 23 are significant, with no significant decrease. Full replication is observed on Zoom and MME-EN at both scales, and additionally on HR-4K and MMStar at 8B. Notably, the $3/3$ replication on 8B MME-EN is obtained despite a mean gain of only 0.65 points, illustrating that paired significance reflects consistent error asymmetry rather than effect magnitude alone. In contrast, the margin over the stronger ZwZ baseline is substantially narrower: only MME-EN at 4B reaches $3/3$ replication, and no 8B benchmark does. The two significant decreases are isolated MME-EN seed outcomes against VisionOPD and ZwZ, rather than a systematic reversal. Overall, the strongest effects concentrate on spatially demanding benchmarks, while the smaller but reproducible gains on MME-EN show that the advantage is not restricted to large aggregate margins. This pattern is consistent with greater headroom in spatial reasoning; the paired analysis itself, however, does not identify which error types are corrected.

Improvements across OPD Formulations. Figure 15 more directly isolates the contribution of spatial calibration. Of the 108 seed-level effects, 99 are positive and 40 survive Holm correction, compared with one significant decrease; 16 of the 36 formulation–benchmark–scale cells replicate in at least two seeds, including eight with full $3/3$ replication. Zoom provides the clearest evidence, reaching $3/3$ under standard OPD at both scales, $2/3$ under 4B EOPD, and $3/3$ under 8B EOPD. Despite smaller absolute gains, MME-CN and MME-EN frequently remain Holm-significant, consistent with the greater precision afforded by their larger evaluation sets and discordant-pair counts. Conversely, TreeB accuracy and VStar improve for every seed, formulation, and model scale, yet no individual comparison survives panel-wise correction; these results establish directional reproducibility, but not conclusive per-seed evidence under the present sample sizes. The only significant decrease occurs for 8B standard OPD on MME-EN at seed 41.

D.3 Qualitative Analysis

Figures 16 and 17 present randomly sampled 8B MixOPD rollouts from VStar and TreeBench. For each student-generated prefix, we visualize the predictive entropy $H(q_t)$ of the privileged teacher distribution $q_t = \pi_{\text{tea}}(\cdot \mid \widetilde{\mathbf{c}}_t)$, together with the model's grounding proposals. Four patterns emerge.

Explicit and inspectable grounding. In the displayed cases, MixOPD pairs correct answers with accurate evidence localization, including examples where the relevant region is extremely small or must be identified through nontrivial spatial relations. Unlike ZwZ, VisionOPD, and ImagineOPD, which internalize privileged evidence implicitly, MixOPD expresses its intermediate visual thoughts through explicit image regions in the rollout. The resulting grounding sequence exposes the model's evidence selection and makes its reasoning trace directly inspectable.

Grounding as question-conditioned visual search. The TreeBench rollouts exhibit structured, question-conditioned visual search. The model successively grounds reference objects, candidate targets, and contextual regions needed to resolve their relations. Some student proposals also differ from the overlaid ground-truth boundaries while enclosing the same task-relevant referent. At a fixed grounding context, this behavior illustrates the distinction formalized by Definition 3.2: multiple coordinate realizations can belong to the same valid evidence class. Proposition 3.3 further shows why coordinate-level matching remains restrictive in such cases: even when the evidence-level distributions agree, their KL retains a within-class residual measuring how probability is allocated among exact coordinate traces.

Localized entropy within grounding spans. The entropy traces exhibit sharp, localized peaks within grounding spans, with uncertainty concentrated at a subset of coordinate-bearing positions. In several examples, positions encoding less-significant coordinate digits have higher entropy than those encoding more-significant digits, suggesting that fine numerical perturbations preserve spatial overlap more readily than coarse changes. Lemma E.19 connects valid-realization dilution to the aggregate conditional-entropy budget across coordinate decisions; the observed coordinate diversity and localized entropy peaks are qualitatively consistent with the same dilution mechanism underlying Lemma 3.6.

Complementary token- and span-level signals. Pronounced teacher-entropy peaks also occur at reasoning positions outside parsed grounding spans. MixOPD leaves these positions at unit weight in Eq. equation 3, because complete-proposal geometry is unavailable. At gated positions, EOPD's additive forward-KL term provides mode-coverage supervision for retained teacher-supported branches that are under-covered by the student, as characterized by Corollary E.25 and Lemma E.26. Outside successfully searched spans, EOPD acts without spatial reweighting; at the coordinate-bearing positions of successfully searched spans, MixOPD multiplies both the PPO-clipped OPD surrogate and the additive forward-KL term by the same detached spatial coefficient. Corollary E.38 formalizes this local composition. Thus, EOPD provides token-level mode-coverage supervision, while MixOPD calibrates the strength of the resulting supervision using complete-proposal geometry. This division of labor provides a direct interpretation of their empirical complementarity in reasoning and grounding.

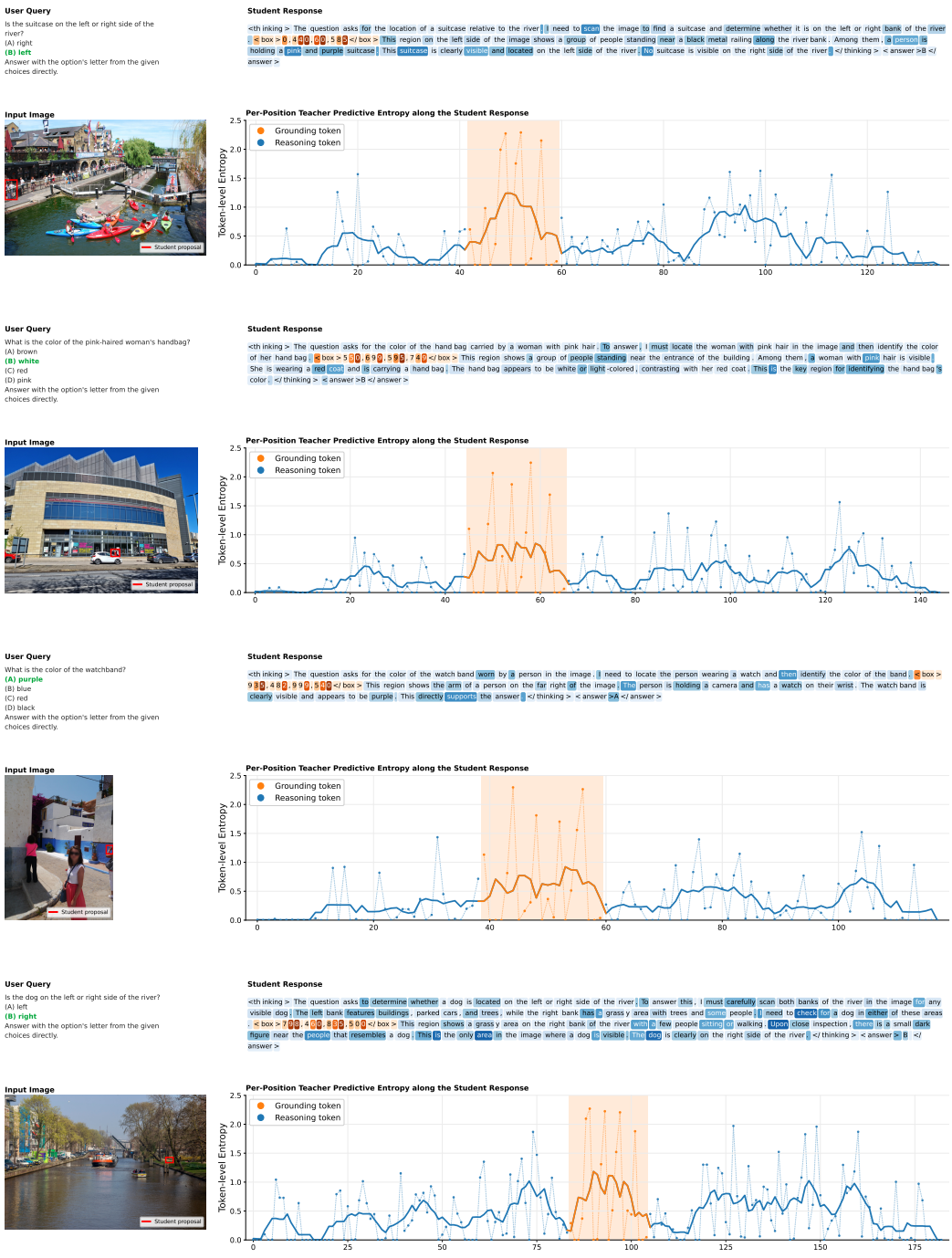

Figure 16: Per-position teacher predictive entropy along the student rollout for representative VStar examples. Specifically, token colors distinguish reasoning (blue) from grounding (orange), with shaded regions marking grounding spans; dotted stems report token-level entropy and solid curves summarize its local trend. Student-proposed boxes are emphasized in *red* for spatial reference.



Figure 17: Per-position teacher predictive entropy along the student rollout for TreeBench examples.

E THEORETICAL ANALYSIS

This appendix separates an analytic population reference from an implementation-faithful analysis of MixOPD. The central question is whether supervision at coordinate-bearing token positions reflects a genuine grounding error or only a different serialization of the same visual evidence. We separate these effects in three steps. First, we construct a normalized probability law over coordinate decisions and decompose its KL through an evidence quotient. Second, we characterize reverse-KL dilution and the mode-coverage role of the forward-KL term added by EOPD. Third, we prove geometry-selective scaling directly for the signed PPO-clipped surrogate used in training. The analysis concerns proposal-local supervision at a fixed student-generated grounding context. The rollout policy, sampled trajectory, teacher distributions, entropy gate, teacher-forced coordinate search, parser, and spatial coefficient are held fixed, so all gradient statements are *local semigradient* statements. At the kinks of the PPO surrogate, the same statements hold for the subgradient selected by automatic differentiation.

Organization. §E.1 defines the coordinate-decision law and the implemented OPD–EOPD surrogate. §E.2 connects that law to evidence classes. §E.3 states the assumptions and clarifies their roles. §E.4 and E.5 analyze dilution and forward-KL mode coverage. §E.6 studies the practical spatial coefficient, §E.7 proves local geometry-selective scaling.

E.1 FORMAL SETUP AND IMPLEMENTED SURROGATES

E.1.1 POPULATION COORDINATE-DECISION LAWS

Let $\mathcal{B}_q$ be the finite space of quantized, positive-area boxes in a normalized image plane. We identify each box with its measurable image region and write

$$d_{\mathrm{J}}(b, b') \coloneqq 1 - \mathrm{IoU}(b, b'), \qquad b, b' \in \mathcal{B}_q,$$

for their Jaccard distance. Fix a student-generated context $c = c_{a_j}$ at the beginning of a grounding span. Let $\tau(b)$ be the canonical token serialization of b, and let $\iota(b)$ be its ordered trace of coordinate-bearing token decisions. Structural tokens (*e.g.*, Tags, separators) are not elements of $\iota(b)$.

Assumption E.1 (Population Canonical Coordinate Grammar and Stopping)**.** *For the ideal population law at every context c, a finite-depth deterministic grammar inserts all structural tokens and exposes only coordinate-bearing decision nodes. At each such node, the policy uses its full-vocabulary next-token kernel evaluated on the candidate-specific canonical prefix. The stopped canonical traces $\{\iota(b) : b \in \mathcal{B}_q\}$ are distinct leaves; every other decision trace is assigned to one absorbing malformed outcome $\perp$.*

This assumption defines an analytic population law over candidate-specific autoregressive traces; it is not a description of the practical teacher-forced search in Appendix B.2. The population intervention fixes only structure: coordinate decisions are neither restricted nor renormalized. A nonnumeric or otherwise invalid decision is therefore absorbed into $\perp$, rather than conditioned away. Variable lengths are padded after stopping, so the resulting decision-tree law is normalized.

Definition E.2 (Raw and Parsed Population Coordinate-Decision Laws)**.** *Let $g_c(y_{<\ell})$ be the deterministic full-token prefix obtained by reinserting the structural grammar around a partial decision trace. For a policy π_u, define the full-vocabulary decision kernel*

$$K^c_{u,\ell}(v \mid y_{<\ell}) \coloneqq \pi_u(v \mid c \oplus g_c(y_{<\ell}))\,,$$

with the absorbing kernel after stopping. Its product defines the raw law over the stopped-trace space $\mathcal{Y}_{c,\mathcal{I}}$, where $L_{c,\mathcal{I}}$ is the padded stopping depth:

$$\mathsf{P}^c_{u,\mathcal{I}}(y) \coloneqq \prod_{\ell=1}^{L_{c,\mathcal{I}}} K^c_{u,\ell}(y_\ell \mid y_{<\ell}).$$

Let

$$\phi_c : \mathcal{Y}_{c,\mathcal{I}} \longrightarrow \Omega_{c,\mathcal{I}} \coloneqq \iota(\mathcal{B}_q) \cup \{\perp\}$$

be the canonical parser. The normalized parsed coordinate-decision law is

$$P^{\mathcal{I}}_u(s \mid c) \coloneqq (\phi_{c\#}\mathsf{P}^c_{u,\mathcal{I}})(s), \qquad s \in \Omega_{c,\mathcal{I}}.$$

For every canonical box,

$$P_u^{\mathcal{I}}(\iota(b) \mid c) = \prod_{\ell=1}^{|\iota(b)|} K_{u,\ell}^{c}(\iota(b)_\ell \mid \iota(b)_{<\ell})\,. \tag{11}$$

Let $\mathsf{G}_c \coloneqq \{y \in \mathcal{Y}_{c,\mathcal{I}} : \phi_c(y) \neq \bot\}$. *If* $P_u^{\mathcal{I}}(\bot \mid c) < 1$, *the induced coordinate law is*

$$P_u^{\mathcal{B}}(b \mid c, \mathsf{G}_c) \coloneqq \frac{P_u^{\mathcal{I}}(\iota(b) \mid c)}{1 - P_u^{\mathcal{I}}(\bot \mid c)}.$$

Lemma E.3 (Normalization and Canonical Atoms)**.** *The raw law* $\mathsf{P}_{u,\mathcal{I}}^{c}$, *its parsed pushforward* $P_u^{\mathcal{I}}$, *and the conditional coordinate law* $P_u^{\mathcal{B}}$ *are normalized. Every canonical trace has a singleton parser fiber.*

Proof. Iterated summation of the full-vocabulary kernels, including the absorbing kernel after stopping, gives $\sum_y \mathsf{P}_{u,\mathcal{I}}^{c}(y) = 1$. A pushforward preserves total mass. Distinct stopped leaves make $\phi_c^{-1}(\iota(b)) = \{\iota(b)\}$, and ordinary conditioning proves the normalization of $P_u^{\mathcal{B}}$. □

For the canonical atom $b = b_j$ of a well-formed realized rollout span, the candidate-specific canonical prefixes in equation 11 are exactly its realized prefixes at the positions $\mathcal{I}(b_j)$. Thus its log-density ratio is the sum of the sampled token log-ratios at precisely the positions weighted by MixOPD. Structural token probabilities are intentionally absent. The outcome $\bot$ absorbs malformed, truncated, out-of-range, and noncanonical traces; it remains part of the evidence-failure class below.

E.1.2 Privileged teacher and implemented local surrogates

For a student rollout $\mathbf{x} \sim \pi_{\theta_{\text{old}}}(\cdot \mid \mathbf{q})$, write

$$c_t = (\mathbf{q}, \mathbf{x}_{<t})$$

for the student context. The teacher may additionally receive visual and logical privileges:

$$\widetilde{c}_t = (c_t, \xi_t^{\text{V}}, \xi_t^{\text{T}}).$$

Below, teacher coordinate laws condition on $\widetilde{c}$; the privilege variables are otherwise suppressed. At a fixed student-generated prefix, define

$$p_t^{\text{old}}(v) \coloneqq \pi_{\theta_{\text{old}}}(v \mid c_t), \qquad p_t^{\theta}(v) \coloneqq \pi_{\theta}(v \mid c_t), \qquad q_t(v) \coloneqq \pi_{\text{tea}}(v \mid \widetilde{c}_t),$$

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

EOPD evaluates its forward-KL term on a truncated teacher distribution. Let $\mathcal{S}_t^L \subseteq \mathcal{V}$ contain the teacher's top-L tokens at position t, let $Z_t^L = q_t(\mathcal{S}_t^L)$, and define

$$\bar{q}_t(v) \coloneqq \frac{q_t(v)\mathbf{1}\{v \in \mathcal{S}_t^L\}}{Z_t^L}.$$

The token truncation level L is distinct from the number K of coordinate candidates used to estimate Spatial Divergence.

Definition E.4 (Implemented OPD–EOPD Family)**.** *Let* $C_A > 0$ *be the sampled log-ratio clip and* $\epsilon_{\text{clip}} \in (0,1)$ *the PPO clip. At a sampled token* x_t, *define*

$$A_t^{\text{raw}} \coloneqq \log q_t(x_t) - \log p_t^{\text{old}}(x_t), \qquad \widehat{A}_t \coloneqq \text{sg}[\text{clip}(A_t^{\text{raw}}, -C_A, C_A)]\,, \tag{12}$$

$$\varrho_t(\theta) \coloneqq \frac{p_t^{\theta}(x_t)}{p_t^{\text{old}}(x_t)}, \tag{13}$$

$$\ell_t^{\text{OPD}}(\theta) \coloneqq \max\big\{-\varrho_t(\theta)\widehat{A}_t, \tag{14}$$

$$-\,\text{clip}(\varrho_t(\theta), 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}})\widehat{A}_t\big\}. \tag{15}$$

Let

$$F_t(\theta) \coloneqq D_{\text{KL}}(\bar{q}_t \| p_t^{\theta}), \qquad g_t \coloneqq \mathbf{1}\{H(q_t) > \eta_H\}.$$

For $\lambda_{\mathrm{F}} \geq 0$, the implemented local surrogate is

$$S_t^{(\lambda_{\mathrm{F}})}(\theta) \coloneqq \ell_t^{\mathrm{OPD}}(\theta) + \lambda_{\mathrm{F}} g_t F_t(\theta). \tag{16}$$

Standard OPD uses $\lambda_{\mathrm{F}} = 0$; EOPD uses $\lambda_{\mathrm{F}} > 0$. Unlike F_t, the OPD term is a signed sampled policy-gradient surrogate, not an analytic KL loss directly backpropagated through the vocabulary distribution.

For distributional analysis only, write

$$R_t(\theta) \coloneqq D_{\mathrm{KL}}(p_t^\theta \| q_t), \qquad \mathcal{J}_t^{(\lambda_{\mathrm{F}})}(\theta) \coloneqq R_t(\theta) + \lambda_{\mathrm{F}} g_t F_t(\theta).$$

At $\theta = \theta_{\mathrm{old}}$, if sampled-log-ratio clipping is inactive on the rollout-policy support, the conditional expected semigradient of ℓ_t^{OPD} equals $\nabla_\theta R_t|_{\theta_{\mathrm{old}}}$. This standard first-update identity explains the reverse-KL interpretation of OPD. We use R_t as a first-update distributional reference and analyze spatial weighting directly through the detached implemented surrogate in Definition E.5. The teacher, $\bar{q}_t$, g_t, and $\widehat{A}_t$ are held fixed. Finite-vocabulary softmax gives the required full support.

Definition E.5 (Detached Coordinate-Span Calibration)**.** *Let $\mathcal{I}(b_j)$ be the coordinate-bearing token positions of the j-th grounding span and let $\widehat{\omega}_j \in [0,1]$ be its estimated spatial coefficient. Define*

$$S_j^{(\lambda_{\mathrm{F}})}(\theta) \coloneqq \sum_{t \in \mathcal{I}(b_j)} S_t^{(\lambda_{\mathrm{F}})}(\theta), \tag{17}$$

$$\widetilde{S}_j^{(\lambda_{\mathrm{F}})}(\theta) \coloneqq \mathrm{sg}(\widehat{\omega}_j) S_j^{(\lambda_{\mathrm{F}})}(\theta), \tag{18}$$

where sg *denotes stop-gradient. Consequently, for a fixed rollout and any selected local gradient or subgradient $G_j \in \partial_\theta S_j^{(\lambda_{\mathrm{F}})}$,*

$$\widetilde{G}_j = \widehat{\omega}_j G_j, \qquad |\widetilde{S}_j^{(\lambda_{\mathrm{F}})}| = \widehat{\omega}_j |S_j^{(\lambda_{\mathrm{F}})}|. \tag{19}$$

Non-grounding and structural positions retain unit weight.

The equality is valid even though $S_j^{(\lambda_{\mathrm{F}})}$ can be signed. EOPD changes the distributional signal by adding a mode-covering component; Spatial Divergence independently changes the local magnitude of the complete implemented signal.

E.2 From Coordinate Decisions to Evidence Classes

E.2.1 Autoregressive and parser decompositions

The grammar intervention above yields an ordinary autoregressive law on a smaller decision tree. Its chain rule therefore contains exactly the coordinate-bearing positions calibrated by MixOPD.

Lemma E.6 (Coordinate-Decision KL Chain Rule)**.** *Let $\mathsf{P}_{\mathcal{I}}$ and $\mathsf{Q}_{\mathcal{I}}$ be two raw stopped decision-trace laws at a fixed grounding context, with full-vocabulary kernels $p_\ell^{\mathcal{I}}$ and $q_\ell^{\mathcal{I}}$. If $\mathsf{P}_{\mathcal{I}} \ll \mathsf{Q}_{\mathcal{I}}$, then*

$$D_{\mathrm{KL}}(\mathsf{P}_{\mathcal{I}} \| \mathsf{Q}_{\mathcal{I}}) = \sum_{\ell=1}^{L_{c,\mathcal{I}}} \mathbb{E}_{Y_{<\ell} \sim \mathsf{P}_{\mathcal{I}}} \left[D_{\mathrm{KL}}\big(p_\ell^{\mathcal{I}}(\cdot \mid Y_{<\ell}) \big\| q_\ell^{\mathcal{I}}(\cdot \mid Y_{<\ell})\big) \right]. \tag{20}$$

If $\mathsf{Q}_{\mathcal{I}} \ll \mathsf{P}_{\mathcal{I}}$, then

$$D_{\mathrm{KL}}(\mathsf{Q}_{\mathcal{I}} \| \mathsf{P}_{\mathcal{I}}) = \sum_{\ell=1}^{L_{c,\mathcal{I}}} \mathbb{E}_{Y_{<\ell} \sim \mathsf{Q}_{\mathcal{I}}} \left[D_{\mathrm{KL}}\big(q_\ell^{\mathcal{I}}(\cdot \mid Y_{<\ell}) \big\| p_\ell^{\mathcal{I}}(\cdot \mid Y_{<\ell})\big) \right]. \tag{21}$$

Let $P^{\mathcal{I}} = \phi_{\#}\mathsf{P}_{\mathcal{I}}$ and $Q^{\mathcal{I}} = \phi_{\#}\mathsf{Q}_{\mathcal{I}}$. Under the first absolute-continuity condition,

$$D_{\mathrm{KL}}(\mathsf{P}_{\mathcal{I}} \| \mathsf{Q}_{\mathcal{I}}) = D_{\mathrm{KL}}(P^{\mathcal{I}} \| Q^{\mathcal{I}}) + \mathbb{E}_{s \sim P^{\mathcal{I}}} D_{\mathrm{KL}}(\mathsf{P}_{\mathcal{I}}(\cdot \mid \phi = s) \| \mathsf{Q}_{\mathcal{I}}(\cdot \mid \phi = s)). \tag{22}$$

Proof. Autoregressive factorization gives

$$\log \frac{\mathsf{P}_{\mathcal{I}}(Y)}{\mathsf{Q}_{\mathcal{I}}(Y)} = \sum_{\ell=1}^{L_{c,\mathcal{I}}} \log \frac{p_\ell^{\mathcal{I}}(Y_\ell \mid Y_{<\ell})}{q_\ell^{\mathcal{I}}(Y_\ell \mid Y_{<\ell})}.$$

Expectation and the tower property prove equation 20; exchanging the laws proves equation 21. For equation 22, factor each raw law into its parsed mass and its conditional law inside the corresponding parser fiber, then average the resulting log-density ratio. □

Remark E.7 (Rollout-Conditioned Local Forward KL)**.** *Equation 21 averages over teacher-generated decision prefixes. EOPD evaluates F_t at student-generated prefixes using the normalized top-L target $\bar{q}_t$. Appendix E.5 therefore analyzes F_t as a rollout-conditioned local forward-KL term.*

E.2.2 Evidence quotient

Fix a targeted visual evidence item $e(c)$. Let $\mathcal{B}_{\mathrm{val}}(c)$ contain the boxes that ground $e(c)$, and let $\mathcal{B}_{\mathrm{inv}}(c) = \mathcal{B}_q \setminus \mathcal{B}_{\mathrm{val}}(c)$. Define

$$\kappa_c : \Omega_{c,\mathcal{I}} \longrightarrow \mathcal{Z}(c) \coloneqq \{e(c), \bot\}$$

by

$$\kappa_c(\iota(b)) = \begin{cases} e(c), & b \in \mathcal{B}_{\mathrm{val}}(c), \\ \bot, & b \in \mathcal{B}_{\mathrm{inv}}(c), \end{cases} \qquad \kappa_c(\bot) = \bot .$$

Thus invalid boxes and malformed traces belong to the same failure fiber. Semantic grounding validity, rather than an IoU threshold, defines the map.

Proposition E.8 (Evidence-Quotient Decomposition)**.** *Let $P^{\mathcal{I}}, Q^{\mathcal{I}}$ be normalized parsed coordinate-decision laws and let*

$$P^{\mathcal{Z}} = \kappa_{c\#} P^{\mathcal{I}}, \qquad Q^{\mathcal{Z}} = \kappa_{c\#} Q^{\mathcal{I}}.$$

If $P^{\mathcal{I}} \ll Q^{\mathcal{I}}$, then

$$D_{\mathrm{KL}}(P^{\mathcal{I}} \| Q^{\mathcal{I}}) = D_{\mathrm{KL}}(P^{\mathcal{Z}} \| Q^{\mathcal{Z}}) + \mathbb{E}_{z \sim P^{\mathcal{Z}}} D_{\mathrm{KL}}\big(P^{\mathcal{I}}(\cdot \mid z) \big\| Q^{\mathcal{I}}(\cdot \mid z)\big) . \tag{23}$$

If also $Q^{\mathcal{I}} \ll P^{\mathcal{I}}$, exchanging the laws gives

$$D_{\mathrm{KL}}(Q^{\mathcal{I}} \| P^{\mathcal{I}}) = D_{\mathrm{KL}}(Q^{\mathcal{Z}} \| P^{\mathcal{Z}}) + \mathbb{E}_{z \sim Q^{\mathcal{Z}}} D_{\mathrm{KL}}\big(Q^{\mathcal{I}}(\cdot \mid z) \big\| P^{\mathcal{I}}(\cdot \mid z)\big) . \tag{24}$$

Proof. For $s \in \Omega_{c,\mathcal{I}}$, let $z = \kappa_c(s)$. Factor

$$\frac{P^{\mathcal{I}}(s)}{Q^{\mathcal{I}}(s)} = \frac{P^{\mathcal{Z}}(z)}{Q^{\mathcal{Z}}(z)} \frac{P^{\mathcal{I}}(s \mid z)}{Q^{\mathcal{I}}(s \mid z)}.$$

Taking logarithms, averaging under $P^{\mathcal{I}}$, and grouping by z proves equation 23; the other direction is identical. □

The first term measures disagreement in grounding validity. The second is a nonnegative within-class residual; on the valid fiber, it measures allocation among distinct coordinate traces that ground the same evidence. Hence a coordinate-token KL can remain positive even after the evidence-level laws agree exactly.

Lemma E.9 (f-Divergence Contraction under the Evidence Map)**.** *Let $f : \mathbb{R}_+ \to \mathbb{R}$ be convex with $f(1) = 0$, and define*

$$D_f(P \| Q) \coloneqq \sum_{s:\, Q(s) > 0} Q(s) f\left(\frac{P(s)}{Q(s)}\right).$$

If $P^{\mathcal{I}} \ll Q^{\mathcal{I}}$, then

$$D_f(P^{\mathcal{I}} \| Q^{\mathcal{I}}) \geq D_f(P^{\mathcal{Z}} \| Q^{\mathcal{Z}}). \tag{25}$$

For strictly convex f, equality holds only when $P^{\mathcal{I}}/Q^{\mathcal{I}}$ is $Q^{\mathcal{I}}$-almost surely constant inside every positive-mass fiber.

Proof. Let $R(s) = P^{\mathcal{I}}(s)/Q^{\mathcal{I}}(s)$. Conditional on a fiber, $P^{\mathcal{Z}}(z)/Q^{\mathcal{Z}}(z) = \mathbb{E}_{Q^{\mathcal{I}}}[R(S) \mid \kappa_c(S) = z]$. Conditional Jensen's inequality proves equation 25; strict convexity gives the equality condition. □

Remark E.10 (Complete two-stage decomposition)**.** *Combining Lemma E.6 and Proposition E.8 decomposes raw coordinate-decision reverse KL into evidence disagreement, the within-class residual, and the nonnegative parser-fiber residual in equation 22. The latter vanishes on every canonical atom because its parser fiber is a singleton, while the aggregated malformed outcome can retain a nonzero residual. Accordingly, the evidence decomposition is stated for the normalized parsed law.*

E.3 Assumptions and Their Roles

The results below use separate assumptions for valid-realization dilution, student exact-trace concentration, teacher-core reliability, and EOPD branch recovery. Because $\mathcal{B}_q$ is finite, every displayed bound is a finite-M statement. Any $\Theta(\log M)$ shorthand refers to a family of contexts or quantizations indexed by increasing M, with the stated constants uniform over that family. This dilution count M is distinct from the practical position-wise partial-value cap M_{c}.

Assumption E.11 (Bounded Valid-Realization Dilution)**.** *At a fixed context c, the teacher assigns total mass $\rho \geq \rho_{\min} > 0$ to $M \geq 2$ distinct valid coordinate traces*

$$\mathcal{A}_M \coloneqq \{\iota(b_{j,1}), \ldots, \iota(b_{j,M})\}, \qquad b_{j,m} \in \mathcal{B}_{\mathrm{val}}(c).$$

Writing

$$q_m \coloneqq P^{\mathcal{I}}_{\mathrm{tea}}(\iota(b_{j,m}) \mid \widetilde{c}), \qquad \sum_{m=1}^{M} q_m = \rho,$$

there exist $\delta \in (0,1)$ and an imbalance factor $1 \leq \Gamma \leq M^{1-\delta}$ such that

$$\frac{\rho}{\Gamma M} \leq q_m \leq \frac{\Gamma \rho}{M}, \qquad m = 1, \ldots, M. \tag{26}$$

The corresponding boxes have Jaccard diameter at most ε.

The factor Γ permits nonuniform teacher allocation while excluding concentration on only a fixed number of traces. The exponent δ measures the effective rate of dilution: the largest normalized atom is at most $M^{-\delta}$.

Assumption E.12 (Student Exact-Trace Concentration Rate)**.** *For the realized valid student proposal $b_j = b_{j,m^\star}$, its coordinate-trace probability under the rollout policy satisfies*

$$p_j \coloneqq P^{\mathcal{I}}_{\theta_{\mathrm{old}}}(\iota(b_j) \mid c) \geq p_0 M^{-\beta}, \qquad p_0 > 0, \quad 0 \leq \beta < \delta. \tag{27}$$

The fixed-$p_{\min}$ condition in the main text is the special case $\beta = 0$. Allowing $\beta > 0$ makes explicit that the pointwise pressure depends on the difference between the teacher's dilution rate and the student's own concentration rate. In a growing problem family, $\rho_{\min}, \delta, p_0, \beta$ are uniform in M; only Γ, the context, and the induced laws may vary.

Assumption E.13 (Reliable Teacher Core)**.** *The population teacher box law $Q^{\mathcal{B}}_c$ admits a nonempty, context-dependent valid core $\varnothing \neq \mathcal{V}_c \subseteq \mathcal{B}_{\mathrm{val}}(c)$ such that*

$$Q^{\mathcal{B}}_c(\mathcal{V}_c) \geq 1 - \nu_c, \qquad \sup_{\substack{b \in \mathcal{B}_{\mathrm{val}}(c) \\ \widetilde{b} \in \mathcal{V}_c}} d_{\mathrm{J}}(b, \widetilde{b}) \leq \varepsilon_c, \tag{28}$$

where $\nu_c, \varepsilon_c \in [0,1]$.

The quantity ν_c is the teacher's off-core mass. It upper-bounds semantically invalid well-formed proposals, but may also include low-mass valid outliers outside $\mathcal{V}_c$. The cross-radius in equation 28 ensures that the high-mass core geometrically covers every valid realization. Malformed probability has already been removed by conditioning on G_c. This assumption concerns the population teacher box law $Q^{\mathcal{B}}_c$. The clean valid-support condition in the main text is instead imposed directly on the practical retained law $\widehat{P}^{\mathcal{B}}_{\mathrm{tea},K}$.

Assumption E.14 (Gate Coverage of a Retained Missing Mode)**.** *At every local position to which the EOPD recovery guarantee is applied, there is a branch set $A_t \subseteq \mathcal{S}^L_t$ such that*

$$g_t = 1, \qquad \bar{q}_t(A_t) - p^{\theta}_t(A_t) \geq \gamma_t > 0. \tag{29}$$

The recovery guarantee therefore requires three conditions at the student prefix: activation of the entropy gate, retention of the relevant teacher-supported branch after Top-L truncation, and student under-coverage of that branch. Appendix E.4 characterizes the entropy signal motivating the gate, while its empirical coverage is assessed separately.

E.4 Reverse KL under Valid-Realization Dilution

Lemma E.15 (Dilution-Induced Exact-Coordinate Pressure)**.** *Under Assumptions E.11 and E.12, define the realized coordinate-trace log-density ratio*

$$r_j^{\mathrm{coord}} := \log \frac{p_j}{q_{m^\star}}.$$

Then

$$\log \frac{Mp_j}{\Gamma\rho} \le r_j^{\mathrm{coord}} \le \log \frac{\Gamma M p_j}{\rho}, \tag{30}$$

and consequently

$$(\delta-\beta)\log M + \log p_0 \le r_j^{\mathrm{coord}} \le (2-\delta)\log M + \log \frac{1}{\rho_{\min}}. \tag{31}$$

For the realized well-formed span,

$$r_j^{\mathrm{coord}} = \sum_{t\in\mathcal{I}(b_j)} \log \frac{p_t^{\mathrm{old}}(x_t)}{q_t(x_t)} = -\sum_{t\in\mathcal{I}(b_j)} A_t^{\mathrm{raw}}. \tag{32}$$

Hence, along any problem family satisfying the assumptions uniformly, $r_j^{\mathrm{coord}} = \Theta(\log M)$.

Proof. The allocation bounds in equation 26 give

$$\frac{Mp_j}{\Gamma\rho} \le \frac{p_j}{q_{m^\star}} \le \frac{\Gamma M p_j}{\rho},$$

proving equation 30. For the lower rate, use $p_j \ge p_0 M^{-\beta}$, $\Gamma \le M^{1-\delta}$, and $\rho \le 1$. For the upper rate, use $p_j \le 1$, the same bound on Γ, and $\rho \ge \rho_{\min}$. Finally, the canonical-atom product in equation 11 contains exactly the realized positions $\mathcal{I}(b_j)$, which proves equation 32. □

Remark E.16 (Pressure, clipping, and the implemented surrogate)**.** *The quantity r_j^{coord} is the unclipped sampled log-density ratio and becomes positive once $M > p_0^{-1/(\delta-\beta)}$. OPD clips each A_t^{raw} before constructing the PPO surrogate, yielding*

$$\bar{r}_j^{\mathrm{coord}} := -\sum_{t\in\mathcal{I}(b_j)} \mathrm{clip}(A_t^{\mathrm{raw}}, -C_A, C_A) \quad \textit{satisfies} \quad |\bar{r}_j^{\mathrm{coord}}| \le C_A |\mathcal{I}(b_j)|.$$

This per-token clipping caps the realized coordinate signal, while PPO ratio clipping determines the update after the initial rollout policy. Lemma E.15 characterizes the pre-clipping serialization pressure; Theorem E.34 directly characterizes the implemented surrogate.

For $a, b \in [0,1]$, write

$$\mathrm{kl}(a\|b) := a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}, \qquad h_2(a) := -a\log a - (1-a)\log(1-a),$$

with the standard continuous extensions at the boundary.

Proposition E.17 (Population Coordinate-RKL Lower Bounds)**.** *Let $P^{\mathcal{I}}$ and $Q^{\mathcal{I}}$ be the student and teacher parsed coordinate-decision laws, with $P^{\mathcal{I}} \ll Q^{\mathcal{I}}$. Under Assumption E.11, the following statements hold.*

(i) For any $s_\star = \iota(b_{j,m^\star})$, let $p = P^{\mathcal{I}}(s_\star)$ and $q = Q^{\mathcal{I}}(s_\star)$. Then

$$D_{\mathrm{KL}}(P^{\mathcal{I}}\|Q^{\mathcal{I}}) \ge \mathrm{kl}(p\|q) \ge p\log\frac{1}{q} - h_2(p). \tag{33}$$

In particular, if $p \ge p_{\min} > 0$, then

$$D_{\mathrm{KL}}(P^{\mathcal{I}}\|Q^{\mathcal{I}}) \ge [p_{\min}\delta\log M - \log 2]_+, \tag{34}$$

where $[a]_+ := \max\{a, 0\}$.

(ii) Let $\sigma = P^{\mathcal{I}}(\mathcal{A}_M) > 0$, and define

$$\bar{p}_m = \frac{P^{\mathcal{I}}(\iota(b_{j,m}))}{\sigma}, \qquad \bar{q}_m = \frac{q_m}{\rho}.$$

For $\sigma, \rho \in (0,1)$,

$$D_{\mathrm{KL}}(P^{\mathcal{I}} \| Q^{\mathcal{I}}) = \mathrm{kl}(\sigma\|\rho) + \sigma D_{\mathrm{KL}}(\bar{p}\|\bar{q}) + (1-\sigma) D_{\mathrm{KL}}\big(P^{\mathcal{I}}(\cdot \mid \mathcal{A}_M^{\mathrm{c}}) \big\| Q^{\mathcal{I}}(\cdot \mid \mathcal{A}_M^{\mathrm{c}})\big). \tag{35}$$

Consequently,

$$D_{\mathrm{KL}}(P^{\mathcal{I}} \| Q^{\mathcal{I}}) \geq \sigma \left[\delta \log M - H(\bar{p})\right]_+. \tag{36}$$

If additionally

$$\sigma \geq \sigma_0 > 0, \qquad H(\bar{p}) \leq \beta_s \log M + C, \qquad \beta_s < \delta,$$

then

$$D_{\mathrm{KL}}(P^{\mathcal{I}} \| Q^{\mathcal{I}}) \geq \sigma_0 \left[(\delta - \beta_s) \log M - C\right]_+. \tag{37}$$

Proof. Data processing through the indicator $\mathbf{1}\{S = s_\star\}$ gives the first inequality in equation 33. Moreover,

$$\mathrm{kl}(p\|q) = -h_2(p) - p \log q - (1-p)\log(1-q) \geq p \log \frac{1}{q} - h_2(p).$$

Since $q \leq \Gamma\rho/M \leq M^{-\delta}$ and $h_2(p) \leq \log 2$, equation 34 follows.

For part (ii), apply the KL chain rule to membership in $\mathcal{A}_M$, which gives equation 35. Because $\bar{q}_m \leq \Gamma/M$,

$$D_{\mathrm{KL}}(\bar{p}\|\bar{q}) = -H(\bar{p}) + \sum_m \bar{p}_m \log \frac{1}{\bar{q}_m} \geq \log \frac{M}{\Gamma} - H(\bar{p}) \geq \delta \log M - H(\bar{p}).$$

Dropping the remaining nonnegative terms proves equation 36 and equation 37. □

Remark E.18 (Pointwise and aggregate rates are different)**.** *Assumption E.12 with $\beta > 0$ is sufficient for the pointwise ratio in Lemma E.15 to grow logarithmically along a problem family. It does not by itself imply an $\Omega(\log M)$ population KL bound: an atom of mass $\Theta(M^{-\beta})$ contributes only $\Theta(M^{-\beta} \log M)$. Aggregate growth requires a fixed atom mass or nonvanishing total student mass with controlled conditional entropy, as stated in Proposition E.17.*

Lemma E.19 (Entropy Signature of Dilution)**.** *Under Assumption E.11, the teacher allocation conditioned on $\mathcal{A}_M$, $\bar{q}_m = q_m/\rho$, satisfies*

$$H(\bar{q}) \geq \log \frac{M}{\Gamma} \geq \delta \log M. \tag{38}$$

The entropy of the parsed coordinate-decision law satisfies

$$H(Q^{\mathcal{I}}) \geq h_2(\rho) + \rho\delta \log M \geq \rho_{\min}\delta \log M. \tag{39}$$

Proof. Because $\|\bar{q}\|_\infty \leq \Gamma/M$, Shannon entropy is at least min-entropy, proving equation 38. Decomposing $Q^{\mathcal{I}}$ over $\{\mathcal{A}_M, \mathcal{A}_M^{\mathrm{c}}\}$ gives

$$H(Q^{\mathcal{I}}) = h_2(\rho) + \rho H(\bar{q}) + (1-\rho) H(Q^{\mathcal{I}}(\cdot \mid \mathcal{A}_M^{\mathrm{c}})),$$

where zero-mass conditional terms are omitted. This proves equation 39. □

Remark E.20 (Aggregate Conditional-Entropy Budget)**.** *Deterministic parsing gives $H(\mathsf{Q}_{\mathcal{I}}) \geq H(Q^{\mathcal{I}})$, while the decision-tree entropy chain rule gives*

$$H(\mathsf{Q}_{\mathcal{I}}) = \sum_{\ell=1}^{L_{c,\mathcal{I}}} \mathbb{E}_{Y_{<\ell} \sim \mathsf{Q}_{\mathcal{I}}} H(q_\ell^{\mathcal{I}}(\cdot \mid Y_{<\ell})).$$

Thus equation 39 lower-bounds the total conditional entropy budget across coordinate decisions. As an aggregate bound, this entropy may concentrate at a subset of coordinate decisions. Gate coverage at student-generated prefixes is specified by Assumption E.14.

E.5 FORWARD KL AND EOPD

We work at one fixed student-generated prefix and omit t when no ambiguity arises. Let $p = p_t^\theta$, $q = q_t$, and $\bar{q} = \bar{q}_t$. The analytic reverse KL $R = D_{\mathrm{KL}}(p\|q)$ is a distributional reference; the implemented reverse branch remains ℓ_t^{OPD}. The forward KL $F = D_{\mathrm{KL}}(\bar{q}\|p)$ is directly included in the additive surrogate.

Lemma E.21 (PPO-Surrogate Bound and Initial RKL Bridge)**.** *Let $d_t = -\widehat{A}_t$, so*

$$d_t = \mathrm{sg}\left[\mathrm{clip}\left(\log \frac{p_t^{\mathrm{old}}(x_t)}{q_t(x_t)}, -C_A, C_A\right)\right].$$

At every differentiable point,

$$\nabla_\theta \ell_t^{\mathrm{OPD}} = \chi_t(\theta) d_t \varrho_t(\theta) \nabla_\theta \log p_t^\theta(x_t), \tag{40}$$

where the selected branch coefficient is

$$\chi_t(\theta) \in \begin{cases} \{0\}, & d_t > 0,\ \varrho_t < 1 - \epsilon_{\mathrm{clip}}, \\ \{0\}, & d_t < 0,\ \varrho_t > 1 + \epsilon_{\mathrm{clip}}, \\ [0,1], & d_t > 0,\ \varrho_t = 1 - \epsilon_{\mathrm{clip}}, \\ [0,1], & d_t < 0,\ \varrho_t = 1 + \epsilon_{\mathrm{clip}}, \\ \{1\}, & \textit{otherwise}. \end{cases}$$

When $d_t = 0$, the gradient is zero irrespective of χ_t. Moreover,

$$\ell_t^{\mathrm{OPD}}(\theta) \geq -(1 + \epsilon_{\mathrm{clip}}) C_A. \tag{41}$$

At $\theta = \theta_{\mathrm{old}}$, if $|\log[p_t^{\mathrm{old}}(x_t)/q_t(x_t)]| < C_A$, then

$$\ell_t^{\mathrm{OPD}}(\theta_{\mathrm{old}}) = \log \frac{p_t^{\mathrm{old}}(x_t)}{q_t(x_t)}. \tag{42}$$

If, more strongly, $|\log[p_t^{\mathrm{old}}(v)/q_t(v)]| < C_A$ for every $v \in \operatorname{supp} p_t^{\mathrm{old}}$, then

$$\mathbb{E}_{X \sim p_t^{\mathrm{old}}}\left[\nabla_\theta \ell_t^{\mathrm{OPD}}(\theta; X)\right]_{\theta=\theta_{\mathrm{old}}} = \nabla_\theta D_{\mathrm{KL}}(p_t^\theta \| q_t)\big|_{\theta=\theta_{\mathrm{old}}}. \tag{43}$$

Proof. Write

$$\ell_t^{\mathrm{OPD}} = \max\{\varrho_t d_t,\ \mathrm{clip}(\varrho_t, 1 - \epsilon_{\mathrm{clip}}, 1 + \epsilon_{\mathrm{clip}}) d_t\}.$$

Differentiating the active branch and using $\nabla_\theta \varrho_t = \varrho_t \nabla_\theta \log p_t^\theta(x_t)$ gives equation 40. If $d_t \geq 0$, the loss is nonnegative. If $d_t < 0$, its clipped branch is at least $(1 + \epsilon_{\mathrm{clip}}) d_t$, proving equation 41. At the rollout policy, $\varrho_t = 1$, so both PPO branches equal d_t; the stated realized condition proves equation 42. Under the support-wise condition,

$$\mathbb{E}_{X \sim p_t^{\mathrm{old}}}\left[\log \frac{p_t^{\mathrm{old}}(X)}{q_t(X)} \nabla_\theta \log p_t^\theta(X)\right]_{\theta=\theta_{\mathrm{old}}}$$

is the softmax score-function expression for the RKL gradient; the derivative of the normalizing $+1$ term has zero expectation. □

Remark E.22. *Equations equation 42–equation 43 characterize the initial rollout policy under their stated clipping conditions. After policy drift, the live ratio and PPO clipping determine the update. Because $\widehat{\omega}_j$ depends on the complete sampled box and can correlate with the coordinate tokens, the implementation-level analysis below uses the detached weighted surrogate in Definition E.5.*

Lemma E.23 (Coarse-Grained Mode-Coverage Asymmetry)**.** *Let $A \subseteq \mathcal{V}$, and define*

$$s = p(A), \qquad r = q(A), \qquad \bar{r} = \bar{q}(A).$$

For $0 < s, r < 1$, KL chain rules over $\{A, A^c\}$ give

$$D_{\mathrm{KL}}(p\|q) = \mathrm{kl}(s\|r) + s D_{\mathrm{KL}}(p_A\|q_A) + (1 - s) D_{\mathrm{KL}}(p_{A^c}\|q_{A^c}), \tag{44}$$

$$D_{\mathrm{KL}}(\bar{q}\|p) = \mathrm{kl}(\bar{r}\|s) + \bar{r} D_{\mathrm{KL}}(\bar{q}_A\|p_A) + (1 - \bar{r}) D_{\mathrm{KL}}(\bar{q}_{A^c}\|p_{A^c}). \tag{45}$$

with zero-mass conditional terms understood by continuity. Consequently,

$$D_{\mathrm{KL}}(p\|q) \geq \mathrm{kl}(s\|r), \tag{46}$$

$$D_{\mathrm{KL}}(\bar{q}\|p) \geq \bar{r}\log\frac{1}{s} - h_2(\bar{r}). \tag{47}$$

For fixed $0 < r < 1$ and $\bar{r} > 0$, $\mathrm{kl}(s\|r) \to -\log(1-r)$ as $s \downarrow 0$, whereas the forward-KL lower bound diverges.

Proof. The decompositions follow by conditioning on the indicator of A. Dropping nonnegative conditional terms proves equation 46. Moreover,

$$\mathrm{kl}(\bar{r}\|s) = -h_2(\bar{r}) + \bar{r}\log\frac{1}{s} + (1-\bar{r})\log\frac{1}{1-s},$$

which proves equation 47. □

Remark E.24. *The distinction is in barrier strength: as $s \downarrow 0$, the coarse-grained reverse-KL barrier approaches the finite value $-\log(1-r)$, whereas the forward-KL lower bound diverges.*

Corollary E.25 (Coverage Barrier of the Implemented Additive Surrogate)**.** *Let $\lambda_{\mathrm{F}} > 0$ and $g_t = 1$. For any retained branch $A \subseteq \mathcal{S}_t^L$, let $s = p_t^\theta(A)$ and $\bar{r} = \bar{q}_t(A) > 0$. Then*

$$S_t^{(\lambda_{\mathrm{F}})}(\theta) \geq -(1+\epsilon_{\mathrm{clip}})C_A + \lambda_{\mathrm{F}}\left[\bar{r}\log\frac{1}{s} - h_2(\bar{r})\right]. \tag{48}$$

Thus the implemented additive surrogate retains a diverging coverage barrier as $s \downarrow 0$, despite its signed PPO component.

Proof. Combine equation 41 and equation 47. □

Lemma E.26 (Missing-Mode Gradients)**.** *Let z be the live student logits and $p = \mathrm{softmax}(z)$. For every token i, the distributional reference and the implemented FKL component satisfy*

$$\frac{\partial R}{\partial z_i} = p_i\left(\log\frac{p_i}{q_i} - R\right), \tag{49}$$

$$\frac{\partial F}{\partial z_i} = p_i - \bar{q}_i. \tag{50}$$

If $i \in \mathcal{S}^L$ and $\bar{q}_i > 0$, then

$$\lim_{p_i \downarrow 0}\frac{\partial F}{\partial z_i} = -\bar{q}_i, \tag{51}$$

while the RKL reference gradient vanishes under full teacher support.

Now add a common scalar a to all logits in a branch $A \subseteq \mathcal{S}^L$, and let $s = p(A)$, $\bar{r} = \bar{q}(A)$. Then

$$\frac{\partial F}{\partial a} = s - \bar{r}, \tag{52}$$

$$G_t^{\mathrm{OPD}}(A) := \frac{\partial \ell_t^{\mathrm{OPD}}}{\partial a} = \chi_t d_t \varrho_t\left[\mathbf{1}\{x_t \in A\} - s\right], \tag{53}$$

$$-\frac{\partial S_t^{(\lambda_{\mathrm{F}})}}{\partial a} = \lambda_{\mathrm{F}} g_t(\bar{r} - s) - G_t^{\mathrm{OPD}}(A). \tag{54}$$

Thus, when $g_t = 1$ and $\bar{r} - s \geq \gamma > 0$, the added FKL component alone contributes a recovery signal of at least $\lambda_{\mathrm{F}}\gamma$.

Proof. Differentiating the two KLs through the softmax gives equation 49 and equation 50. Under full teacher support, $R \leq \log(1/q_{\min})$, where $q_{\min} = \min_{v\in\mathcal{V}} q(v) > 0$, and $p_i|\log p_i| \to 0$, giving the limits. A common branch shift satisfies $\partial \log p(x_t)/\partial a = \mathbf{1}\{x_t \in A\} - p(A)$. Combine this identity with equation 40 to obtain equation 53; the remaining equations follow directly. □

Remark E.27 (Total-Surrogate Recovery under Noncancellation). *The added FKL component yields the recovery signal in equation 52. The complete additive surrogate satisfies a corresponding bound under the following noncancellation condition:*

$$G_t^{\mathrm{OPD}}(A) \le (1-\kappa)\lambda_{\mathrm{F}}(\bar{r}-s), \qquad \kappa \in (0,1], \tag{55}$$

which gives

$$-\frac{\partial S_t^{(\lambda_{\mathrm{F}})}}{\partial a} \ge \kappa\lambda_{\mathrm{F}}(\bar{r}-s).$$

Spatial weighting multiplies both components by the same nonnegative detached coefficient and preserves their within-position balance.

E.6 Spatial Geometry and Practical Approximation

E.6.1 Population spatial coefficient

Definition E.28 (Spatial Cross-Discrepancy). *For $P, Q \in \mathcal{P}(\mathcal{B}_q)$, define*

$$\mathcal{D}_{\mathrm{spa}}(P,Q) \coloneqq \mathbb{E}_{\substack{b\sim P \\ b'\sim Q}}\left[d_{\mathrm{J}}(b,b')\right].$$

For a realized student proposal b, write

$$\omega_Q(b) \coloneqq \mathcal{D}_{\mathrm{spa}}(\delta_b, Q) = \mathbb{E}_{b'\sim Q}[d_{\mathrm{J}}(b,b')]\,. \tag{56}$$

Lemma E.29 (Basic Spatial Properties). *For all $P, Q \in \mathcal{P}(\mathcal{B}_q)$ and $b, \widetilde{b} \in \mathcal{B}_q$,*

$$0 \le \mathcal{D}_{\mathrm{spa}}(P,Q) \le 1, \tag{57}$$

$$\mathcal{D}_{\mathrm{spa}}(P,Q) = \mathcal{D}_{\mathrm{spa}}(Q,P), \tag{58}$$

$$\left|\omega_Q(b) - \omega_Q(\widetilde{b})\right| \le d_{\mathrm{J}}(b,\widetilde{b}). \tag{59}$$

If total variation is defined as

$$d_{\mathrm{TV}}(P,Q) \coloneqq \sup_A |P(A)-Q(A)| = \tfrac{1}{2}\|P-Q\|_1,$$

then

$$|\omega_P(b) - \omega_Q(b)| \le d_{\mathrm{TV}}(P,Q). \tag{60}$$

Proof. Boundedness and symmetry follow from the corresponding properties of d_{J}. The Jaccard distance is a metric on positive-area measurable regions, so the reverse triangle inequality gives

$$|d_{\mathrm{J}}(b,b') - d_{\mathrm{J}}(\widetilde{b},b')| \le d_{\mathrm{J}}(b,\widetilde{b}).$$

Taking expectation proves equation 59. Equation 60 follows because $d_{\mathrm{J}}(b,\cdot) \in [0,1]$. □

Lemma E.30 (Teacher-Relative Spatial Sandwich). *Let $Q \in \mathcal{P}(\mathcal{B}_q)$ have nonempty finite support $S = \mathrm{supp}(Q)$. Define*

$$d_S(b) \coloneqq \min_{\widetilde{b}\in S} d_{\mathrm{J}}(b,\widetilde{b}), \qquad \varepsilon_S \coloneqq \max_{\widetilde{b},\widetilde{b}'\in S} d_{\mathrm{J}}(\widetilde{b},\widetilde{b}').$$

Then, for every $b \in \mathcal{B}_q$,

$$d_S(b) \le \omega_Q(b) \le \min\{1, d_S(b) + \varepsilon_S\}. \tag{61}$$

Proof. The lower bound follows because $d_{\mathrm{J}}(b,\widetilde{b}) \ge d_S(b)$ for every $\widetilde{b} \in S$. Let $b^\star \in S$ attain the minimum. By the triangle inequality,

$$d_{\mathrm{J}}(b,\widetilde{b}) \le d_{\mathrm{J}}(b,b^\star) + d_{\mathrm{J}}(b^\star,\widetilde{b}) \le d_S(b) + \varepsilon_S.$$

Take expectation under Q and use the unit upper bound. □

The sandwich shows that the spatial coefficient is a smoothed distance to the teacher support. If $b \in S$, then $\omega_Q(b) \le \varepsilon_S$; if $d_S(b) \ge \Delta$, then $\omega_Q(b) \ge \Delta$.

E.6.2 TEACHER-FORCED CONSTRAINED COORDINATE SEARCH AND TRUNCATION

Let the ideal autoregressive population teacher coordinate law be

$$Q_c^{\mathcal{B}} := P_{\mathrm{tea}}^{\mathcal{B}}(\cdot \mid \widetilde{c}, \mathsf{G}_c).$$

Fix a well-formed realized span b_j, with coordinate lengths $\{\ell_{j,h}\}_{h=1}^4$ and digit positions $\{t_{j,h,r}\}$, and suppress the realized span template from the notation. The teacher-forcing pass provides

$$q_{j,h,r}(v) := \pi_{\mathrm{tea}}(v \mid \widetilde{c}_{t_{j,h,r}}), \qquad \mathcal{D}_{j,h,r} := \mathcal{D}_{\mathrm{dig}} \cap \mathrm{TopV}(q_{j,h,r}, V),$$

where $\mathcal{D}_{\mathrm{dig}}$ contains the ten single-digit tokens and the Top-V operation is over the full vocabulary. The realized structural tokens and coordinate lengths are held fixed. Any box matching this template is assigned the factorized score

$$u_j(b) := \sum_{h=1}^{4} \sum_{r=1}^{\ell_{j,h}} \log q_{j,h,r}(d_{h,r}(b)). \tag{62}$$

The search uses only digits satisfying $d_{h,r}(b) \in \mathcal{D}_{j,h,r}$. Within each coordinate, it extends the partial-value list one digit at a time and retains its Top-M_{c} entries after every extension. Tie-aware best-first Cartesian search then applies the admissibility predicate $z_2 - z_1 \geq s_{\min} = 16$ in each ordered axis-pair stage before cutoff K. Inadmissible pairs are skipped without counting toward K or the cutoff-tie group; out-of-range admitted pairs are subsequently removed. If either range-filtered constrained axis list is empty, the search returns $(\varnothing, \varnothing, 1)$, and no practical candidate law is constructed. Otherwise, the final merge uses cutoff K, retains the complete boundary tie group, and applies the fixed tie-break before selecting at most K boxes. Every retained box therefore satisfies the minimum-side constraint. On a successful search, the resulting nonempty candidate set is

$$\mathcal{A}_{j,V,M_{\mathrm{c}},K}(c) := \mathcal{B}_{\mathrm{tea},j}^{K} \subseteq \mathcal{B}_q.$$

We fix $s_{\min} = 16$ and suppress it, together with the fixed argument c, from the notation where unambiguous. This construction uses only the cached realized-prefix teacher laws; it does not decode or rescore candidates on their own prefixes. If $b = b_j$, then

$$u_j(b_j) = \log P_{\mathrm{tea}}^{\mathcal{I}}(\iota(b_j) \mid \widetilde{c}),$$

because the cached contexts coincide with the realized prefixes. For an alternative candidate, later factors remain conditioned on the realized student prefixes; hence the practical law need not equal the population law conditioned on its retained set. If $\mathcal{A}_{j,V,M_{\mathrm{c}},K}(c) \neq \varnothing$, define its retained population mass

$$\alpha_{j,V,M_{\mathrm{c}},K}(c) := Q_c^{\mathcal{B}}(\mathcal{A}_{j,V,M_{\mathrm{c}},K}(c)) > 0$$

and the exact conditional law

$$Q_{c,A_j}^{\mathcal{B}} := Q_c^{\mathcal{B}}(\cdot \mid \mathcal{A}_{j,V,M_{\mathrm{c}},K}(c)).$$

The implementation constructs a practical law $\widehat{Q}_{j,V,M_{\mathrm{c}},K}^{\mathcal{B}}$, denoted by $\widehat{q}_j$ in Algorithm 1, by normalizing the factorized scores in equation 62:

$$\widehat{Q}_{j,V,M_{\mathrm{c}},K}^{\mathcal{B}}(b) := \frac{\exp u_j(b)}{\sum_{\widetilde{b} \in \mathcal{A}_{j,V,M_{\mathrm{c}},K}(c)} \exp u_j(\widetilde{b})}, \qquad b \in \mathcal{A}_{j,V,M_{\mathrm{c}},K}(c).$$

It is zero outside the retained set. Its frozen-context factorization gap relative to the conditional population law is

$$\zeta_{j,V,M_{\mathrm{c}},K} := d_{\mathrm{TV}}\left(\widehat{Q}_{j,V,M_{\mathrm{c}},K}^{\mathcal{B}}, Q_{c,A_j}^{\mathcal{B}}\right).$$

This gap need not vanish; no smallness assumption is imposed. For later use, define

$$\widehat{\omega}_{j,V,M_{\mathrm{c}},K}(b) := \omega_{\widehat{Q}_{j,V,M_{\mathrm{c}},K}^{\mathcal{B}}}(b).$$

Lemma E.31 (Finite-Candidate Approximation)**.** *For every nonempty candidate set with* $\alpha_{j,V,M_{\mathrm{c}},K} > 0$*, we have*

$$d_{\mathrm{TV}}\left(Q_{c,A_j}^{\mathcal{B}}, Q_c^{\mathcal{B}}\right) = 1 - \alpha_{j,V,M_{\mathrm{c}},K}, \tag{63}$$

$$\xi_{j,V,M_{\mathrm{c}},K} := d_{\mathrm{TV}}\left(\widehat{Q}_{j,V,M_{\mathrm{c}},K}^{\mathcal{B}}, Q_c^{\mathcal{B}}\right) \leq \min\{1, 1 - \alpha_{j,V,M_{\mathrm{c}},K} + \zeta_{j,V,M_{\mathrm{c}},K}\}, \tag{64}$$

$$\left|\widehat{\omega}_{j,V,M_{\mathrm{c}},K}(b) - \omega_{Q_c^{\mathcal{B}}}(b)\right| \leq \xi_{j,V,M_{\mathrm{c}},K}. \tag{65}$$

Under Assumption E.13, let

$$\eta_{j,V,M_c,K} \coloneqq \widehat{Q}^{\mathcal{B}}_{j,V,M_c,K}(\mathcal{V}_c^c)$$

be the off-core mass of the practical law. Then

$$\eta_{j,V,M_c,K} \le \min\left\{1,\ \nu_c + \xi_{j,V,M_c,K},\ \frac{\nu_c}{\alpha_{j,V,M_c,K}} + \zeta_{j,V,M_c,K}\right\}. \tag{66}$$

Proof. For brevity, let $A = \mathcal{A}_{j,V,M_c,K}(c)$, $\alpha = \alpha_{j,V,M_c,K}$, $\widehat{Q} = \widehat{Q}^{\mathcal{B}}_{j,V,M_c,K}$, $\xi = \xi_{j,V,M_c,K}$, and $\zeta = \zeta_{j,V,M_c,K}$. Then

$$Q^{\mathcal{B}}_{c,A_j}(b) = \frac{Q^{\mathcal{B}}_c(b)\mathbf{1}\{b \in A\}}{\alpha}.$$

Directly summing the positive and negative parts of $Q^{\mathcal{B}}_{c,A_j} - Q^{\mathcal{B}}_c$ gives $d_{\mathrm{TV}}(Q^{\mathcal{B}}_{c,A_j}, Q^{\mathcal{B}}_c) = 1 - \alpha$. The triangle inequality proves equation 64, and Lemma E.29 gives equation 65.

For the first contamination bound, total variation transfers event probabilities:

$$\widehat{Q}(\mathcal{V}_c^c) \le Q^{\mathcal{B}}_c(\mathcal{V}_c^c) + \xi \le \nu_c + \xi.$$

For the retained-law bound,

$$Q^{\mathcal{B}}_{c,A_j}(\mathcal{V}_c^c) = \frac{Q^{\mathcal{B}}_c(A \cap \mathcal{V}_c^c)}{\alpha} \le \frac{\nu_c}{\alpha}.$$

Adding the factorization-gap event-probability error ζ proves equation 66. □

Remark E.32 (Candidate-Set Controls and Approximation Quality)**.** *The integers V, M_c, K determine the candidate set. Here V controls the full-vocabulary shortlist before digit filtering, M_c caps each partial coordinate-value list after every digit extension, and K sets the cutoff for both axis-pair stages and the final merge while bounding the final retained set. Approximation quality is governed jointly by retained mass, candidate validity, and the factorization gap. Non-nested candidate sets can yield nonmonotonic approximation as V, M_c, or K increases. If $\nu_c = 0$ and $\zeta_{j,V,M_c,K} = 0$, the practical retained law is fully supported on the valid core, so $\eta_{j,V,M_c,K} = 0$ even when $\alpha_{j,V,M_c,K}$ is small. Truncation can then alter the population expectation while preserving valid–invalid geometric selectivity. More generally, the retained-mass certificate is informative when $\nu_c/\alpha_{j,V,M_c,K} + \zeta_{j,V,M_c,K} < 1$; with $\zeta_{j,V,M_c,K} = 0$, this reduces to $\alpha_{j,V,M_c,K} > \nu_c$.*

Hence, whenever the practical candidate law is defined, Lemma E.33 below applies with $\widehat{Q} = \widehat{Q}^{\mathcal{B}}_{j,V,M_c,K}$, $\eta = \eta_{j,V,M_c,K}$, and $\varepsilon = \varepsilon_c$.

Lemma E.33 (Robust Spatial Selectivity)**.** *Let $\widehat{Q} \in \mathcal{P}(\mathcal{B}_q)$ satisfy*

$$\widehat{Q}(\mathcal{V}_c) \ge 1 - \eta, \qquad \varnothing \ne \mathcal{V}_c \subseteq \mathcal{B}_{\mathrm{val}}(c), \qquad \sup_{\substack{b \in \mathcal{B}_{\mathrm{val}}(c) \\ \widetilde{b} \in \mathcal{V}_c}} d_{\mathrm{J}}(b, \widetilde{b}) \le \varepsilon,$$

for $\eta, \varepsilon \in [0,1]$, and define

$$\widehat{\omega}(b) \coloneqq \mathbb{E}_{\widetilde{b} \sim \widehat{Q}}\left[d_{\mathrm{J}}(b, \widetilde{b})\right].$$

Fix $\Delta \in [0,1]$, and set

$$\varepsilon_{\mathrm{eff}} \coloneqq (1-\eta)\varepsilon + \eta, \qquad \Delta_{\mathrm{eff}} \coloneqq (1-\eta)\Delta. \tag{67}$$

Then:

(i) if $b \in \mathcal{B}_{\mathrm{val}}(c)$, then

$$\widehat{\omega}(b) \le \varepsilon_{\mathrm{eff}};$$

(ii) if $b \in \mathcal{B}_{\mathrm{inv}}(c)$ and

$$d_{\mathrm{val}}(b; c) \coloneqq \inf_{\widetilde{b} \in \mathcal{B}_{\mathrm{val}}(c)} d_{\mathrm{J}}(b, \widetilde{b}) \ge \Delta,$$

then

$$\widehat{\omega}(b) \ge \Delta_{\mathrm{eff}}.$$

Proof. For $b \in \mathcal{B}_{\mathrm{val}}(c)$, split the expectation over $\mathcal{V}_c$ and $\mathcal{V}_c^c$:

$$\widehat{\omega}(b) \le \widehat{Q}(\mathcal{V}_c)\varepsilon + \widehat{Q}(\mathcal{V}_c^c) = \varepsilon + \widehat{Q}(\mathcal{V}_c^c)(1-\varepsilon) \le (1-\eta)\varepsilon + \eta.$$

If $d_{\mathrm{val}}(b;c) \ge \Delta$, every core element is valid, so the contribution from $\mathcal{V}_c$ alone gives

$$\widehat{\omega}(b) \ge \widehat{Q}(\mathcal{V}_c)\Delta \ge (1-\eta)\Delta.$$

□

E.7 Unified Geometry-Selective Guarantee

We now state the implementation-level result. It applies to standard OPD and EOPD with arbitrary sampled-log-ratio and PPO clipping because it is proved directly for $S_j^{(\lambda_{\mathrm{F}})}$, not for an analytic KL proxy.

Theorem E.34 (Implementation-Faithful Geometry-Selective Scaling)**.** *Fix a rollout, a grounding context c, and a successful, well-formed grounding span j. Let $\widehat{Q}$ be its practical teacher box law satisfying the hypotheses of Lemma E.33 for a fixed $\Delta \in [0,1]$. Set*

$$\widehat{\omega}_j := \mathbb{E}_{\widetilde{b}\sim\widehat{Q}}\left[d_{\mathrm{J}}(b_j,\widetilde{b})\right].$$

Let $S_j^{(\lambda_{\mathrm{F}})}(\theta)$ and $\widetilde{S}_j^{(\lambda_{\mathrm{F}})}(\theta)$ be defined in Definition E.5. For any selected local gradient

$$G_j = \nabla_\theta S_j^{(\lambda_{\mathrm{F}})},$$

or

$$G_j \in \partial_{\mathrm{C}} S_j^{(\lambda_{\mathrm{F}})}$$

at a PPO kink, let $\widetilde{G}_j$ be its calibrated counterpart. Then

$$\widetilde{G}_j = \widehat{\omega}_j G_j,$$

and

$$\left|\widetilde{S}_j^{(\lambda_{\mathrm{F}})}\right| = \widehat{\omega}_j \left|S_j^{(\lambda_{\mathrm{F}})}\right|, \qquad \left\|\widetilde{G}_j\right\| = \widehat{\omega}_j \left\|G_j\right\|.$$

(i) ***Valid-proposal suppression.*** *If $b_j \in \mathcal{B}_{\mathrm{val}}(c)$, then*

$$\left|\widetilde{S}_j^{(\lambda_{\mathrm{F}})}\right| \le \varepsilon_{\mathrm{eff}} \left|S_j^{(\lambda_{\mathrm{F}})}\right|, \tag{68}$$

$$\left\|\widetilde{G}_j\right\| \le \varepsilon_{\mathrm{eff}} \left\|G_j\right\|. \tag{69}$$

(ii) ***Separated-proposal signal retention.*** *If $b_j \in \mathcal{B}_{\mathrm{inv}}(c)$ and $d_{\mathrm{val}}(b_j;c) \ge \Delta$, then*

$$\left|\widetilde{S}_j^{(\lambda_{\mathrm{F}})}\right| \ge \Delta_{\mathrm{eff}} \left|S_j^{(\lambda_{\mathrm{F}})}\right|, \tag{70}$$

$$\left\|\widetilde{G}_j\right\| \ge \Delta_{\mathrm{eff}} \left\|G_j\right\|. \tag{71}$$

(iii) ***Aggregate separated-to-valid magnitude selectivity.*** *For this part, let $\mu(\cdot \mid c)$ be any conditional law over successful, well-formed realized grounding spans. For μ-almost every span j, condition on its realized rollout and all detached teacher- and search-side quantities, let $\widehat{Q}_j$ denote its practical teacher box law, and define*

$$\widehat{\omega}_j := \mathbb{E}_{\widetilde{b}\sim\widehat{Q}_j}\left[d_{\mathrm{J}}(b_j,\widetilde{b})\right].$$

Assume that the hypotheses of Lemma E.33 hold μ-almost surely with common parameters $\eta, \varepsilon, \Delta$, and hence with the common effective constants

$$\varepsilon_{\mathrm{eff}} = (1-\eta)\varepsilon + \eta, \qquad \Delta_{\mathrm{eff}} = (1-\eta)\Delta.$$

For either choice

$$h_j = \left|S_j^{(\lambda_F)}\right| \qquad \text{or} \qquad h_j = \|G_j\|,$$

assume that h_j is μ-integrable, and define

$$N \coloneqq \mathbb{E}_\mu\left[h_j \mathbf{1}\left\{b_j \in \mathcal{B}_{\text{val}}(c)\right\}\right], \tag{72}$$

$$S^\Delta \coloneqq \mathbb{E}_\mu\left[h_j \mathbf{1}\left\{b_j \in \mathcal{B}_{\text{inv}}(c),\, d_{\text{val}}(b_j; c) \geq \Delta\right\}\right], \tag{73}$$

$$\widetilde{N} \coloneqq \mathbb{E}_\mu\left[\widehat{\omega}_j h_j \mathbf{1}\left\{b_j \in \mathcal{B}_{\text{val}}(c)\right\}\right], \tag{74}$$

$$\widetilde{S}^\Delta \coloneqq \mathbb{E}_\mu\left[\widehat{\omega}_j h_j \mathbf{1}\left\{b_j \in \mathcal{B}_{\text{inv}}(c),\, d_{\text{val}}(b_j; c) \geq \Delta\right\}\right]. \tag{75}$$

Then

$$\widetilde{N} \leq \varepsilon_{\text{eff}} N, \qquad \widetilde{S}^\Delta \geq \Delta_{\text{eff}} S^\Delta. \tag{76}$$

Whenever $N > 0$, $\widetilde{N} > 0$, and $\varepsilon_{\text{eff}} > 0$,

$$\frac{\widetilde{S}^\Delta}{\widetilde{N}} \geq \frac{\Delta_{\text{eff}}}{\varepsilon_{\text{eff}}} \frac{S^\Delta}{N}. \tag{77}$$

Proof. Under the semigradient convention, $\widehat{\omega}_j$ is a fixed nonnegative scalar. Away from PPO kinks, ordinary differentiation therefore gives

$$\nabla_\theta \widetilde{S}_j^{(\lambda_F)} = \widehat{\omega}_j \nabla_\theta S_j^{(\lambda_F)}.$$

At a PPO kink, positive homogeneity of the Clarke subdifferential gives the corresponding identity

$$\widetilde{G}_j = \widehat{\omega}_j G_j.$$

Because $\widehat{\omega}_j \geq 0$, it also follows directly that

$$\left|\widetilde{S}_j^{(\lambda_F)}\right| = \widehat{\omega}_j \left|S_j^{(\lambda_F)}\right|, \qquad \left\|\widetilde{G}_j\right\| = \widehat{\omega}_j \|G_j\|.$$

Lemma E.33 gives

$$\widehat{\omega}_j \leq \varepsilon_{\text{eff}}$$

for a valid proposal and

$$\widehat{\omega}_j \geq \Delta_{\text{eff}}$$

for an invalid proposal satisfying $d_{\text{val}}(b_j; c) \geq \Delta$. Multiplying these coefficient bounds by the exact magnitude identities proves parts (i)–(ii).

For part (iii), the common-parameter assumption ensures that the same pointwise bounds hold for μ-almost every span. Since h_j is nonnegative and μ-integrable, integrating the valid-proposal bound gives

$$\widetilde{N} \leq \varepsilon_{\text{eff}} N,$$

while integrating the separated-proposal bound gives

$$\widetilde{S}^\Delta \geq \Delta_{\text{eff}} S^\Delta.$$

This proves Equation 76. Finally, whenever $N > 0$, $\widetilde{N} > 0$, and $\varepsilon_{\text{eff}} > 0$,

$$\frac{\widetilde{S}^\Delta}{\widetilde{N}} \geq \frac{\Delta_{\text{eff}} S^\Delta}{\widetilde{N}} \geq \frac{\Delta_{\text{eff}} S^\Delta}{\varepsilon_{\text{eff}} N} = \frac{\Delta_{\text{eff}}}{\varepsilon_{\text{eff}}} \frac{S^\Delta}{N},$$

which proves Equation 77. □

Remark E.35 (Proposal-Local Interpretation)**.** *For each realized span, the base surrogate determines the local update direction, while Spatial Divergence determines the retained magnitude through $\widehat{\omega}_j$. Parts (i)–(ii) give the corresponding pointwise bounds for valid and separated proposals, and part (iii) integrates the same bounds over a conditional law of successful spans.*

Remark E.36 (When the magnitude ratio is amplified)**.** *Provided $S^\Delta > 0$ and the denominators are positive, Equation 77 certifies strict amplification when*

$$\Delta_{\text{eff}} > \varepsilon_{\text{eff}} \quad \Longleftrightarrow \quad (1 - \eta)(\Delta - \varepsilon) > \eta.$$

If $\varepsilon_{\text{eff}} = 0$, the valid-proposal budget is eliminated and the ratio is interpreted without division. A nontrivial separated-proposal certificate additionally requires $\eta < 1$.

Corollary E.37 (Initial Unclipped Bridge to Dilution Pressure)**.** *Under the conditions of Theorem E.34 and Assumptions E.11 and E.12, suppose* $b_j = b_{j,m^\star} \in \mathcal{B}_{\mathrm{val}}(c)$ *and*

$$M > p_0^{-1/(\delta-\beta)}.$$

Define the auxiliary calibrated ratio

$$r_j^{\mathrm{Mix}} := \widehat{\omega}_j r_j^{\mathrm{coord}}.$$

Then

$$0 \le r_j^{\mathrm{Mix}} \le \varepsilon_{\mathrm{eff}} r_j^{\mathrm{coord}}, \tag{78}$$

$$r_j^{\mathrm{coord}} - r_j^{\mathrm{Mix}} \ge (1-\varepsilon_{\mathrm{eff}})\left[(\delta-\beta)\log M + \log p_0\right]. \tag{79}$$

If, additionally, $\theta = \theta_{\mathrm{old}}$ *and sampled-log-ratio clipping is inactive at every* $t \in \mathcal{I}(b_j)$*, then*

$$S_j^{(0)}(\theta_{\mathrm{old}}) = r_j^{\mathrm{coord}}, \qquad \widetilde{S}_j^{(0)}(\theta_{\mathrm{old}}) = r_j^{\mathrm{Mix}}. \tag{80}$$

With clipping or subsequent PPO updates, the exact statements for the implemented surrogate are instead those of Theorem E.34.

Proof. Lemma E.15 makes $r_j^{\mathrm{coord}} > 0$, while robust valid-proposal selectivity gives $\widehat{\omega}_j \le \varepsilon_{\mathrm{eff}}$. This proves the first two inequalities. Summing equation 42 over $\mathcal{I}(b_j)$ and applying equation 32 proves equation 80. □

Corollary E.38 (Geometry-Calibrated EOPD)**.** *Under the conditions of Theorem E.34, let* $\lambda_{\mathrm{F}} > 0$*. Fix a gated position* $t \in \mathcal{I}(b_j)$*, so* $g_t = 1$*, and a branch* $A_t \subseteq \mathcal{S}_t^L$ *satisfying*

$$\bar{q}_t(A_t) - p_t^\theta(A_t) \ge \gamma_t > 0.$$

For parts (ii)–(iv)*, further suppose that* $b_j \in \mathcal{B}_{\mathrm{inv}}(c)$ *and* $d_{\mathrm{val}}(b_j; c) \ge \Delta$*.*

(i) The value-magnitude and subgradient-norm bounds in Theorem E.34 apply to the complete implemented additive surrogate and separately to its OPD and FKL components.

(ii) Under a common logit shift a *on* A_t*, the calibrated FKL component obeys*

$$-\frac{\partial}{\partial a}\left[\mathrm{sg}(\widehat{\omega}_j)\lambda_{\mathrm{F}} F_t\right] = \widehat{\omega}_j \lambda_{\mathrm{F}}\left[\bar{q}_t(A_t) - p_t^\theta(A_t)\right]. \tag{81}$$

Moreover,

$$-\frac{\partial}{\partial a}\left[\mathrm{sg}(\widehat{\omega}_j)\lambda_{\mathrm{F}} F_t\right] \ge \Delta_{\mathrm{eff}}\lambda_{\mathrm{F}}\gamma_t. \tag{82}$$

(iii) Let $s = p_t^\theta(A_t)$*,* $\bar{r} = \bar{q}_t(A_t)$*, and suppose the actual sampled OPD branch satisfies*

$$G_t^{\mathrm{OPD}}(A_t) \le (1-\kappa)\lambda_{\mathrm{F}}(\bar{r} - s)$$

for some $\kappa \in (0, 1]$*. For the same separated invalid proposal,*

$$-\frac{\partial}{\partial a}\left[\mathrm{sg}(\widehat{\omega}_j) S_t^{(\lambda_{\mathrm{F}})}\right] \ge \Delta_{\mathrm{eff}}\kappa\lambda_{\mathrm{F}}\gamma_t. \tag{83}$$

(iv) At the same gated position and separated invalid proposal, the value of the calibrated local surrogate satisfies

$$\mathrm{sg}(\widehat{\omega}_j) S_t^{(\lambda_{\mathrm{F}})} \ge -(1+\epsilon_{\mathrm{clip}})C_A + \Delta_{\mathrm{eff}}\lambda_{\mathrm{F}}\left[\bar{r}\log\frac{1}{s} - h_2(\bar{r})\right]_+. \tag{84}$$

Thus any $\Delta_{\mathrm{eff}} > 0$ *preserves a diverging missing-mode barrier.*

Proof. Part (i) applies the theorem to the sum and to either component. Part (ii) uses equation 52, stop-gradient, and $\widehat{\omega}_j \ge \Delta_{\mathrm{eff}}$. Part (iii) combines equation 54 with the stated noncancellation condition and the same spatial lower bound. For part (iv), $\widehat{\omega}_j \ell_t^{\mathrm{OPD}} \ge -(1+\epsilon_{\mathrm{clip}})C_A$ because $\widehat{\omega}_j \in [0, 1]$, while $F_t \ge [\bar{r}\log(1/s) - h_2(\bar{r})]_+$; apply the spatial lower bound to the nonnegative FKL term. □

Remark E.39 (Complementarity). *EOPD and MixOPD address different failures. The added FKL component protects retained teacher-supported modes, while the sampled OPD component retains serialization-sensitive supervision. Spatial calibration multiplies both components by the same coefficient, preserving their within-span balance while suppressing their local magnitudes on valid proposals and retaining a controlled fraction on separated errors. Their empirical gains may therefore compose.*

Corollary E.40 (Recovery of the Main-Text Guarantee). *Under the conditions of Theorem E.34, take the idealized valid-support setting*

$$\mathcal{V}_c = \mathcal{B}_{\mathrm{val}}(c), \qquad \eta = 0,$$

and suppose the valid set has Jaccard diameter at most ε. Then $\varepsilon_{\mathrm{eff}} = \varepsilon$ and $\Delta_{\mathrm{eff}} = \Delta$. Hence the implemented coordinate-position surrogate magnitude and local subgradient norm are scaled by at most ε on a valid proposal and by at least Δ on an invalid proposal separated by Δ. If, in addition, Assumptions E.11 and E.12 hold with $\beta = 0$, $M > p_0^{-1/\delta}$, $\theta = \theta_{\mathrm{old}}$, and sampled-log-ratio clipping is inactive on $\mathcal{I}(b_j)$, the same bound suppresses the positive raw dilution ratio in the implemented span surrogate.

Proof. Substitute the ideal conditions into equation 67, then apply Theorem E.34. Under the additional conditions, apply Corollary E.37. □

Table 13: Notation used throughout the paper. Boldface on contexts and fixed context, span, or search indices are suppressed when unambiguous; bound indices and one-use dummy variables are omitted.

Notation	Definition
Rollouts, policies, contexts, and conventions	
$\mathcal{Q}, \mathbf{q}$	Training-query distribution (or dataset) and a sampled image–text query.
$\mathbf{x} = (x_1, \ldots, x_T),\ x_{<t}$	Student rollout of length T, its token at position t, and the strict prefix preceding that position.
$\pi_\theta, \pi_{\theta_{\mathrm{old}}}, \pi_{\mathrm{tea}}$	Live student, detached rollout policy $\pi_{\theta_{\mathrm{old}}} = \mathrm{sg}(\pi_\theta)$, and fixed privileged teacher.
$\mathbf{c}_t, \widetilde{\mathbf{c}}_t$	Student context $\mathbf{c}_t = (\mathbf{q}, \mathbf{x}_{<t})$ and privileged teacher context. The appendices write $c_t, \widetilde{c}_t$, and the local analysis fixes $c = c_{a_j}$.
$\xi_t^{\mathrm{V}}, \xi_t^{\mathrm{T}}$	Visual and logical teacher privileges, respectively, with $\widetilde{c}_t = (c_t, \xi_t^{\mathrm{V}}, \xi_t^{\mathrm{T}})$.
$\mathcal{V}$	Shared token vocabulary; distinct from the valid teacher box core $\mathcal{V}_c$.
$\mathcal{M}, m_t$	Policy-update minibatch and binary valid-response-token mask. All OPD-based objectives use the global valid-token denominator $\sum_{\mathcal{M}} m_t$.
$\mathrm{sg}, \mathbf{1}, \mathrm{clip}$	Stop-gradient, indicator, and interval-clipping operators.
$\mathcal{P}(\mathcal{X}),\ f_{\#}P,\ P \ll Q$	Probability laws on $\mathcal{X}$, pushforward of P through f, and absolute continuity.
Grounding spans, coordinate grammar, and induced laws	
j, J, a_j, b_j	Grounding-span index, number of parsed well-formed spans, start position of span j, and its realized student box proposal.
$\mathcal{J}_{\mathrm{suc}}$	Indices of well-formed student grounding spans for which practical candidate search succeeds; their coordinate-position sets are pairwise disjoint.
$\mathcal{B}$ (main), $\mathcal{B}_q$ (appendix)	Finite space of canonical quantized positive-area boxes in the normalized image plane.
$\tau(b), \iota(b), \mathcal{I}(b_j)$	Canonical full-token serialization of b, its ordered stopped trace of coordinate-bearing decisions, and the corresponding rollout positions. Structural tokens belong to $\tau(b)$ but not to $\iota(b)$ or $\mathcal{I}(b_j)$.
$g_c(y_{<\ell}), K^c_{u,\ell}$	Deterministic grammar completion of a partial decision trace and the full-vocabulary kernel induced by policy π_u at coordinate position ℓ.
$\mathcal{Y}_{c,\mathcal{I}}, L_{c,\mathcal{I}}$	Raw stopped-trace space and padded stopping depth. The main text abbreviates the latter as $L_{\mathcal{I}}$.

Continued on next page

Table 13 (continued)

Notation	Definition
$\mathsf{P}^{c}_{u,\mathcal{I}}$	Normalized raw stopped coordinate-decision law obtained by multiplying $K^{c}_{u,\ell}$ through the padded depth.
$\phi_c, \Omega_{c,\mathcal{I}}, \perp$	Canonical parser, parsed outcome space $\Omega_{c,\mathcal{I}} = \iota(\mathcal{B}_q) \cup \{\perp\}$, and absorbing malformed-trace atom. The main text writes Ω_c.
$P^{\mathcal{I}}_u$	Normalized parsed coordinate-decision law $\phi_{c\#}\mathsf{P}^{c}_{u,\mathcal{I}}$, retaining malformed mass at $\perp$.
$P^{\mathcal{I}}_{\mathrm{tea}}, P^{\mathcal{I}}_{\theta_{\mathrm{old}}}$	Teacher and rollout-policy instances of $P^{\mathcal{I}}_u$, conditioned on the privileged and student grounding contexts, respectively.
$\mathsf{G}_c, P^{\mathcal{B}}_u$	Event that a raw trace is well formed and the induced box law obtained by conditioning $P^{\mathcal{I}}_u$ on G_c.
$\mathsf{P}_{\mathcal{I}}, \mathsf{Q}_{\mathcal{I}}, p^{\mathcal{I}}_{\ell}, q^{\mathcal{I}}_{\ell}, Y$	Generic pair of raw stopped-trace laws, their decision kernels, and a raw trace random variable in the autoregressive KL chain rules.
$P^{\mathcal{I}}, Q^{\mathcal{I}}$	Generic pair of normalized parsed coordinate-decision laws; policy subscripts are suppressed in distributional propositions.
Evidence quotient and information quantities	
$e(c), \mathcal{B}_{\mathrm{val}}(c), \mathcal{B}_{\mathrm{inv}}(c)$	Target visual evidence, boxes that validly ground it, and the complementary set of semantically invalid well-formed boxes.
$\varepsilon(c)$	Jaccard diameter of $\mathcal{B}_{\mathrm{val}}(c)$; abbreviated as ε at a fixed context.
$\kappa_c, \mathcal{Z}(c), \perp$	Evidence map and quotient space $\mathcal{Z}(c) = \{e(c), \perp\}$. The atom $\perp$ collects invalid boxes and $\perp$, whereas $\perp$ denotes parser-level malformation only.
$P^{\mathcal{Z}}_u, P^{\mathcal{Z}}, Q^{\mathcal{Z}}$	Evidence-level pushforwards of parsed coordinate laws through κ_c.
$D_{\mathrm{KL}}, \mathcal{D}_{\mathrm{RKL}}, \mathcal{D}_{\mathrm{FKL}}$	KL divergence and the orientations $\mathcal{D}_{\mathrm{RKL}}(p,q) = D_{\mathrm{KL}}(p\|q)$ and $\mathcal{D}_{\mathrm{FKL}}(q,p) = D_{\mathrm{KL}}(q\|p)$.
$H(P)$	Shannon entropy of a discrete law P.
$D_f(P\|Q), f$	Generic f-divergence and its convex generator, with $f(1) = 0$, used for contraction under the evidence map.
$\mathrm{kl}(a\|b), h_2(a), [a]_+$	Bernoulli KL, binary entropy, and positive part $\max\{a, 0\}$.
Implemented OPD–EOPD objectives	
$p^{\mathrm{old}}_t, p^{\theta}_t, q_t$	Rollout-student, live-student, and teacher next-token laws. The main-text shorthand p_t denotes p^{θ}_t.
$A^{\mathrm{raw}}_t, \widehat{A}_t, C_A$	Sampled teacher–rollout log ratio $A^{\mathrm{raw}}_t = \log q_t(x_t) - \log p^{\mathrm{old}}_t(x_t)$, its detached clipped version, and the symmetric clipping radius.
$\varrho_t(\theta), \epsilon_{\mathrm{clip}}$	PPO probability ratio $\varrho_t(\theta) = p^{\theta}_t(x_t)/p^{\mathrm{old}}_t(x_t)$ at the sampled token x_t, and the corresponding PPO clipping radius.
$\ell^{\mathrm{OPD}}_t(\theta)$	Implemented signed PPO-clipped per-token OPD surrogate.
$\mathcal{S}^L_t, Z^L_t, \bar{q}_t$	Teacher Top-L token set, retained mass $Z^L_t = q_t(\mathcal{S}^L_t)$, and renormalized truncated teacher law. The main text writes $\bar{q}^L_t$.
η_H, g_t	Teacher-entropy threshold and detached gate $g_t = \mathbf{1}\{H(q_t) > \eta_H\}$.
$R_t(\theta), F_t(\theta)$	Analytic reverse-KL reference $D_{\mathrm{KL}}(p^{\theta}_t\|q_t)$ and implemented forward-KL term $D_{\mathrm{KL}}(\bar{q}_t\|p^{\theta}_t)$.
$\lambda_{\mathrm{F}}, \mathcal{J}^{(\lambda_{\mathrm{F}})}_t$	Nonnegative FKL weight and analytic reference $R_t + \lambda_{\mathrm{F}} g_t F_t$; $\lambda_{\mathrm{F}} = 0$ recovers standard OPD and $\lambda_{\mathrm{F}} > 0$ gives EOPD.
$S^{(\lambda_{\mathrm{F}})}_t(\theta)$	Implemented local surrogate $\ell^{\mathrm{OPD}}_t + \lambda_{\mathrm{F}} g_t F_t$.
w_t	Detached token weight: $\widehat{\omega}_{j,K}$ on $t \in \mathcal{I}(b_j)$ for a successfully searched span, and 1 otherwise. Search failure recovers the base objective.
$\mathcal{L}_{\mathrm{OPD}}, \mathcal{L}_{\mathrm{EOPD}}, \mathcal{L}^{(\lambda_{\mathrm{F}})}_{\mathrm{MixOPD}}$	Globally valid-token-averaged standard OPD, additive EOPD, and spatially calibrated objectives.
$S^{(\lambda_{\mathrm{F}})}_j, \widetilde{S}^{(\lambda_{\mathrm{F}})}_j$	Base coordinate-span surrogate $\sum_{t \in \mathcal{I}(b_j)} S^{(\lambda_{\mathrm{F}})}_t$ and its calibrated form $\mathrm{sg}(\widehat{\omega}_j) S^{(\lambda_{\mathrm{F}})}_j$.
$G_j, \widetilde{G}_j, \partial_{\mathrm{C}}$	Selected base and calibrated local (sub)gradients; ∂_{C} is the Clarke subdifferential at PPO kinks.
$d_t, \chi_t(\theta)$	Auxiliary PPO quantities $d_t = -\widehat{A}_t$ and the selected branch coefficient $\chi_t \in [0, 1]$ in the local derivative.

Continued on next page

Table 13 (continued)

Notation	Definition
GRPO and DAPO baseline objectives	
$G, \mathbf{x}_i, T_i$	Number of responses sampled per query, response i, and its response length in the group-normalized RL baselines.
$R_{\mathrm{acc},i}, R_{\mathrm{dIoU},i}, R_i$	Binary answer-accuracy reward, dual-IoU grounding reward, and their sum $R_i = R_{\mathrm{acc},i} + R_{\mathrm{dIoU},i}$.
$\widehat{A}_i, \epsilon_{\mathrm{adv}}, \varrho_{i,t}(\theta)$	Detached group-normalized response advantage, its numerical stabilizer, and the live-to-rollout likelihood ratio for token t of response i.
$\ell_{i,t}^{\mathrm{clip}}(\theta; \epsilon_-, \epsilon_+), \ell_{i,t}^{(c)}$	Asymmetrically PPO-clipped token loss and its VeRL dual-clipped form with negative-advantage threshold c.
$C_{\mathrm{GRPO}}, C_{\mathrm{DAPO}}, \epsilon_{\mathrm{low}}, \epsilon_{\mathrm{high}}$	Method-specific dual-clip thresholds and DAPO's asymmetric lower and upper PPO clipping radii.
$\widehat{K}_{i,t}^{\mathrm{ref}}, \beta$	VeRL low-variance $K3$ reference-KL estimator and its GRPO weight. This use of β is distinct from the student concentration exponent below.
$\mathcal{L}_{\mathrm{GRPO}}, \mathcal{L}_{\mathrm{DAPO}}$	Implemented GRPO and DAPO objectives defined in Appendix B.3.
Spatial supervision and annotation-based diagnostics	
$\mathrm{IoU}, d_{\mathrm{J}}$	Box intersection-over-union and the Jaccard distance $d_{\mathrm{J}}(b, b') = 1 - \mathrm{IoU}(b, b')$ used in the formal analysis.
$\mathcal{D}_{\mathrm{spa}}(P, Q)$	Expected Jaccard distance under independent $b \sim P, b' \sim Q$; a bounded spatial cross-discrepancy, not a proper statistical divergence.
$\delta_b, \omega_Q(b)$	Dirac law at b and population spatial coefficient $\omega_Q(b) = \mathcal{D}_{\mathrm{spa}}(\delta_b, Q)$. The Dirac subscript is unrelated to the dilution exponent δ.
$\mathcal{B}_j^f, \mathcal{B}_j^{\mathrm{exp}}, \mathcal{B}_j^{\mathrm{MAP}}, \mathcal{B}_j^{\mathrm{uni}}$	Generic span-level spatial supervision law and its probability-weighted teacher, MAP point-mass, and support-uniform instantiations in §3.2.
$\widehat{\omega}_{j,K}^{\mathrm{exp}}, \widehat{\omega}_{j,K}^{\mathrm{mode}}, \widehat{\omega}_{j,K}^{\mathrm{unif}}$	Practical expectation-, MAP-, and support-uniform spatial coefficients; $\widehat{\omega}_{j,K}^{\mathrm{exp}} \equiv \widehat{\omega}_{j,K}$ is used by MixOPD.
$\widehat{b}_j^{\mathrm{MAP}}, u$	Highest-probability retained teacher proposal and expected spatial agreement $u = 1 - \widehat{\omega}_{j,K}$ in the pressure-field analysis.
$r_j^{\mathrm{coord}}, \bar{r}_j^{\mathrm{coord}}, r_j^{\mathrm{Mix}}$	Unclipped realized coordinate log-density ratio, its advantage-clipped diagnostic sum, and the auxiliary calibrated ratio $r_j^{\mathrm{Mix}} = \widehat{\omega}_j r_j^{\mathrm{coord}}$.
$\mathcal{B}_j^\star, K_j, \{b_{j,k}\}_{k=1}^{K_j}$	Post-hoc matched reference-annotation set, actual retained support size, and the retained teacher proposals for an eligible span; eligible spans satisfy $K_j \geq 2$.
$U_{j,K}^{\mathcal{B}}, U_j, m_j$	Uniform law over the retained boxes, uniform law over retained indices $[K_j]$, and deterministic MAP index satisfying $b_{j,m_j} = \widehat{b}_j^{\mathrm{MAP}}$.
$q_{j,k}, r_{j,k}, \bar{r}_j$	Teacher mass of proposal $b_{j,k}$, its annotation-relative Jaccard error, and uniform candidate risk $\bar{r}_j = \mathbb{E}_{k\sim U_j}[r_{j,k}]$.
$S_j^{\mathrm{geo}}, S_j^{\mathrm{prob}}$	Geometric and probabilistic selectivity, with $S_j^{\mathrm{geo}} = \mathrm{Std}_{k\sim U_j}[r_{j,k}]$ and $S_j^{\mathrm{prob}} = \frac{K_j}{\sqrt{K_j - 1}} \mathrm{Std}_{k\sim U_j}[q_{j,k}]$.
$\rho_j^{\mathrm{align}}, A_j^{\mathrm{align}}, A_j^{\mathrm{MAP}}$	Probability–geometry alignment, soft aligned-selection gain, and hard MAP-selection gain, where $\rho_j^{\mathrm{align}} = -\mathrm{Corr}_{k\sim U_j}(q_{j,k}, r_{j,k})$, $A_j^{\mathrm{align}} = \mathbb{E}_{k\sim U_j}[r_{j,k}] - \mathbb{E}_{k\sim q_j}[r_{j,k}]$, and $A_j^{\mathrm{MAP}} = \bar{r}_j - r_{j,m_j}$. For nondegenerate spans, $A_j^{\mathrm{align}} = \sqrt{K_j - 1}\, S_j^{\mathrm{geo}} S_j^{\mathrm{prob}} \rho_j^{\mathrm{align}}$.
$d_j^{\mathrm{stu}}, d_j^{\mathrm{tea}}$	Annotation-relative Jaccard errors of the realized student proposal and the retained teacher MAP proposal.
e_j^g, δ_j^f	Absolute calibration error $e_j^g = \lvert\widehat{\omega}_{j,K}^g - d_j^{\mathrm{stu}}\rvert$, $g \in \{\mathrm{exp}, \mathrm{mode}, \mathrm{unif}\}$, and excess error $\delta_j^f = e_j^f - e_j^{\mathrm{exp}}$, $f \in \{\mathrm{mode}, \mathrm{unif}\}$.
$\widehat{q}_j^{\mathrm{TFCS}}, \widehat{q}_j^{\mathrm{AR}}, \widehat{\zeta}_j$	Normalized same-support box laws under factorized TFCS and candidate-specific autoregressive rescoring, and their total-variation gap $\widehat{\zeta}_j = d_{\mathrm{TV}}(\widehat{q}_j^{\mathrm{TFCS}}, \widehat{q}_j^{\mathrm{AR}})$.
$\widehat{\omega}_j^{\mathrm{TFCS}}, \widehat{\omega}_j^{\mathrm{AR}}, \widehat{\omega}_j^{\mathrm{beam}}, \Delta\omega$	Spatial coefficients from TFCS, same-support autoregressive rescoring, and constrained beam search; $\Delta\omega$ denotes their paired difference in the search-fidelity analysis.

Continued on next page

Table 13 (continued)

Notation	Definition
Teacher-forced constrained coordinate search	
$V, M_{\mathrm{c}}, K, s_{\min}$	Full-vocabulary teacher shortlist size, per-position partial-coordinate cap, axis/final merge cutoff (*i.e.*, retained box count), and minimum box side.
$c/h, r, \ell_{j,h}, t_{j,h,r}$	Coordinate index (written c in Appendix B.2 and h in the formal search and complexity analyses), digit index, digit length of coordinate h in span j, and the corresponding rollout position.
$\mathcal{D}_{\mathrm{dig}}, q_{j,h,r}, \mathcal{D}_{j,h,r}$	Ten single-digit tokens, cached full-vocabulary teacher law at digit position $t_{j,h,r}$, and the digit tokens retained from its Top-V list.
$d_{h,r}(b), \widehat{\mathcal{V}}$	Digit token of candidate box b at coordinate h, digit position r, and the Top-V token subset illustrated in Figure 3.
$u_j(b)$	Factorized candidate score obtained by summing cached digit-token log probabilities over the four coordinates; denoted u_b locally in Algo 1.
$Z_{j,h}, \widetilde{X}_j, \widetilde{Y}_j, X_j, Y_j, \widetilde{\mathcal{C}}_j, \mathcal{C}_j$	Retained scored coordinate lists, pre- and post-range-filtered axis-pair lists, and pre- and post-tie-broken final scored-box lists in Algorithm 1.
$\mathcal{R}_j$	Search record formed from the realized box, its digit positions and lengths, and cached teacher laws.
$\mathcal{B}^K_{\mathrm{tea}}(\widetilde{c}_{a_j}), \mathcal{B}^K_{\mathrm{tea},j}, \mathcal{A}_{j,V,M_{\mathrm{c}},K}(c)$	Main-text, algorithmic, and fully indexed theoretical notation for the same retained nonempty teacher candidate set on a successful search.
$\widehat{P}^{\mathcal{B}}_{\mathrm{tea},K}, \widehat{q}_j, \widehat{Q}^{\mathcal{B}}_{j,V,M_{\mathrm{c}},K}$	Main-text, algorithmic, and fully indexed theoretical notation for the normalized practical teacher law obtained from the factorized scores.
$Q^{\mathcal{B}}_c, P^{\mathcal{B}}_{\mathrm{tea}}$	Ideal population teacher box law $Q^{\mathcal{B}}_c = P^{\mathcal{B}}_{\mathrm{tea}}(\cdot \mid \widetilde{c}, \mathsf{G}_c)$ and its underlying teacher box-law notation.
$\alpha_{j,V,M_{\mathrm{c}},K}, Q^{\mathcal{B}}_{c,\mathcal{A}_j}$	Population teacher mass retained by the practical candidate set and the exact population law conditioned on that set.
$\zeta_{j,V,M_{\mathrm{c}},K}, \xi_{j,V,M_{\mathrm{c}},K}$	Frozen-context factorization gap relative to the exact retained conditional law and total practical-to-population search error, both in total variation.
$\widehat{\omega}_{j,K}, \widehat{\omega}_{j,V,M_{\mathrm{c}},K}(b), \widehat{\omega}_j$	Main-text, fully indexed, and locally index-suppressed forms of the detached practical spatial coefficient. Search failure sets $\widehat{\omega}_{j,K} = 1$.
Finite-search error and robust spatial selectivity	
$d_{\mathrm{TV}}(P, Q)$	Total-variation distance $\sup_A \lvert P(A) - Q(A)\rvert = \frac{1}{2}\lVert P - Q\rVert_1$.
$S, d_S(b), \varepsilon_S$	Support $S = \mathrm{supp}(Q)$, distance from b to S, and Jaccard diameter of S in the teacher-relative spatial sandwich.
$\mathcal{V}_c, \nu_c, \varepsilon_c$	High-mass valid teacher box core, population mass outside it, and maximum cross-distance from any valid proposal to the core.
$\eta_{j,V,M_{\mathrm{c}},K}, \eta$	Practical law's mass outside $\mathcal{V}_c$; η is the same quantity after suppressing search indices.
$d_{\mathrm{val}}(b; c), \Delta$	Minimum Jaccard distance from b to $\mathcal{B}_{\mathrm{val}}(c)$ and the required separation of an invalid proposal.
$\varepsilon_{\mathrm{eff}}, \Delta_{\mathrm{eff}}$	Robust valid-proposal upper factor $(1-\eta)\varepsilon + \eta$ and separated-error lower factor $(1-\eta)\Delta$, where ε is the common core cross-radius (equal to the valid-set diameter in the clean main-text setting).
Valid-realization dilution and mode coverage	
$\mathcal{A}_M, M, b_{j,m}, m^\star$	Selected set of $M \geq 2$ valid coordinate traces, their corresponding boxes, and the index realized by the student.
$q_m, \rho, \rho_{\min}$	Teacher atom mass $q_m = P^{\mathcal{I}}_{\mathrm{tea}}(\iota(b_{j,m}) \mid \widetilde{c})$, total mass $\rho = \sum_m q_m$ on $\mathcal{A}_M$, and its uniform positive lower bound.
Γ, δ	Imbalance factor and dilution exponent in $\rho/(\Gamma M) \leq q_m \leq \Gamma\rho/M$, with $1 \leq \Gamma \leq M^{1-\delta}$.
$p_j, p_0, \beta, p_{\min}$	Rollout atom mass $p_j = P^{\mathcal{I}}_{\theta_{\mathrm{old}}}(\iota(b_j) \mid c)$ and its lower-bound parameters $p_j \geq p_0 M^{-\beta}$; the main-text condition $p_j \geq p_{\min}$ is the case $\beta = 0$.
$s_\star, p, q$	Selected canonical coordinate atom and its student and teacher masses in the pointwise population-RKL bound.
$\sigma, \bar{p}_m, \bar{q}_m$	Student mass $\sigma = P^{\mathcal{I}}(\mathcal{A}_M)$ and conditional student/teacher allocations on $\mathcal{A}_M$, with $\bar{q}_m = q_m/\rho$.
σ_0, β_s, C	Auxiliary lower-mass, conditional-entropy-rate, and additive constants in the aggregate population-RKL bound.

Continued on next page

Table 13 (continued)

Notation	Definition
A_t, γ_t	Retained Top-L teacher-supported branch and its under-coverage margin $\bar{q}_t(A_t) - p_t^{\theta}(A_t) \geq \gamma_t > 0$.
$A, s, r, \bar{r}, p_A, q_A, \bar{q}_A$	Generic token branch, its masses under $p, q, \bar{q}$, and the corresponding conditional laws in the coarse-grained KL decompositions.
$z = (z_i)_i, p_i, q_i, \bar{q}_i, a$	Student logits and token probabilities; a is a common logit shift applied to every token in branch A.
$G_t^{\mathrm{OPD}}(A), \kappa, q_{\min}$	Sampled OPD derivative along the branch shift, noncancellation fraction $\kappa \in (0, 1]$ (distinct from κ_c), and minimum teacher token probability used in the missing-mode limit.
Local and aggregate geometry-selective guarantees	
$\widehat{Q}, \widehat{Q}_j$	Locally index-suppressed practical teacher box law for one span and its span-indexed counterpart in the aggregate result.
$\mu(\cdot \mid c), h_j$	Conditional law over successful, well-formed realized spans and their local magnitude, where h_j denotes either $\lvert S_j^{(\lambda_{\mathrm{F}})}\rvert$ or $\lVert G_j\rVert$.
$N, S^{\Delta}, \widetilde{N}, \widetilde{S}^{\Delta}$	Aggregate valid-proposal and Δ-separated-error magnitudes before and after spatial calibration.
Computational analysis	
$B, S_i, \widetilde{S}_i, T_i$	Retained global rollout-batch size and the student-prefix, privileged-teacher-prefix, and response lengths of example i.
N_s, N_t, N_r	Aggregate student-side, teacher-side, and response token counts: $N_s = \sum_i (S_i + T_i)$, $N_t = \sum_i (\widetilde{S}_i + T_i)$, and $N_r = \sum_i T_i$.
$P_s, P_t, E_{\mathrm{PPO}}, N_{\mathrm{mb}}$	Student and teacher parameter counts, number of policy-update epochs, and number of policy minibatches per epoch.
F_{OPD}	Parameter-multiplication FLOP estimate for one retained OPD rollout–update cycle.
$\widehat{k}_{1,t}$	Sampled $K1$ reverse-KL estimator $\log p_t^{\mathrm{old}}(x_t) - \log q_t(x_t) = -A_t^{\mathrm{raw}}$ used in the complexity accounting.
U, N_H	Vocabulary size $U = \lvert\mathcal{V}\rvert$ and number of valid positions whose entropy gate is active.
$J_{\mathrm{tot}}, L_j, L_{\mathrm{coord}}$	Total parsed grounding spans in the batch, number $L_j = \lvert\mathcal{I}(b_j)\rvert$ of coordinate decisions in span j, and total coordinate-bearing token count $L_{\mathrm{coord}} = \sum_j L_j$.
$d_{j,h,r}, d_{\max}, E_{j,h,r}, n_{j,h,r}$	Number of retained digits at one search position, its upper bound $d_{\max} = \min\{V, 10\}$, number of partial-value extensions, and retained partial-list size.
$H_{j,x}, H_{j,y}, H_{j,b}, R_{j,x}, R_{j,y}$	Numbers of states examined by the two axis merges and final box merge, and the returned x- and y-axis list sizes.
$C_{\mathrm{enum}}, C_{\mathrm{merge}}, C_{\mathrm{spa}}$	Coordinate-enumeration, Cartesian-merge, and spatial-normalization, IoU-reduction, and broadcast costs.
$C_{\mathrm{MixOPD}}^{\mathrm{std}}, C_{\mathrm{MixOPD}}^{\mathrm{EOPD}}$	Total per-cycle compute for standard-OPD and EOPD instantiations of MixOPD.
$\bar{N}, N_{\max}, \bar{J}$	Common mean full-sequence length, maximum-length token envelope, and mean number of grounding spans per rollout in the reported scale estimates.
Paired significance analysis	
A_i, B_i, n_{ab}	Binary correctness indicators of MixOPD and its comparator on example i, and paired cell count $n_{ab} = \sum_i \mathbf{1}\{A_i = a, B_i = b\}$.
N, m, H_0	Number of paired examples, number $m = n_{10} + n_{01}$ of discordant pairs, and the McNemar marginal-homogeneity null hypothesis.
$p_{\mathrm{McN}}, \Delta\,\mathrm{Acc}, p_{\mathrm{Holm}}^{\mathrm{adj}}$	Exact two-sided McNemar p-value, paired accuracy difference in percentage points, and panel-wise Holm-adjusted p-value.

F STATEMENTS

F.1 LIMITATIONS AND FUTURE WORK

Our theory is intentionally local. Lemma E.15 diagnoses serialization-sensitive coordinate pressure before clipping, while Theorem E.34 establishes geometry-selective scaling of the implemented local (sub)gradient for a fixed rollout. It does not characterize global training dynamics, and the finite-candidate analysis controls search and reweighting error conditional on candidate coverage, not coverage itself. These boundaries expose a broader question: *what should count as the same evidence?* MixOPD uses Jaccard geometry as a tractable proxy, yet task-equivalent evidence may be spatially disjoint, while overlapping regions need not play the same functional role. Its scalar coefficient also determines how strongly OPD learns from a realization, rather than which geometric correction to make. A natural next step is to learn task-conditioned equivalence over richer evidence structures (*e.g.*, masks, temporal trajectories, and 3D regions) and convert it into directional, structured credit assignment. This could move grounded distillation beyond exact token agreement and fixed spatial primitives while preserving dense supervision and inference-time simplicity. Validation beyond Qwen3-VL and box grounding remains future work.

F.2 BROADER IMPACTS

Our study of spatially calibrated on-policy distillation provides broader insights into the development of grounded multimodal reasoning systems. By using span-level spatial geometry to calibrate the magnitude of coordinate-level learning signals, MixOPD encourages LVLMs to expose task-relevant regions alongside their predictions, empirically improving fine-grained reasoning and making intermediate evidence directly inspectable without adding a teacher, search procedure, parameters, or model forward pass at inference. More broadly, this direction may inspire learning objectives that respect spatial equivalence rather than exact token realizations, advancing more interpretable and evidence-grounded models. Nevertheless, localized evidence does not guarantee causal faithfulness or factual correctness, and inherited data biases remain important considerations in deployment.

F.3 AI USE STATEMENT

Artificial Intelligence (AI) has become increasingly influential in scientific research, but we maintain that creativity, judgment, and responsibility remain fundamentally human. Accordingly, all core scientific contributions of this work, *e.g.*, research ideas, methodological design, theoretical development, experimental design, analysis, interpretation, and conclusions, were conceived and developed by the authors. AI assistants were used only in an assistive role. Specifically, large language models (LLMs) were employed to refine the writing and presentation of the manuscript, including language polishing, sentence rephrasing, readability improvement, and grammar checking, and to assist in verifying the correctness and internal consistency of theoretical derivations. Coding agents were also used during implementation to assist with code debugging. These tools were not used to originate scientific ideas, formulate research claims, determine methodological choices, interpret experimental results, or draw scientific conclusions. All AI-assisted outputs were treated solely as suggestions and independently reviewed by at least two authors before incorporation. All scientific claims, theoretical arguments, experimental results, and final manuscript content underwent human verification, and the authors take full responsibility for the correctness and integrity of this work.