DOPD: DUAL ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) offers superior capacity transfer by supervising student-sampled trajectories with dense token-level signals. To furnish high-quality supervision sources and thereby elevate the performance frontier of distillation, an intuitive direction is to infuse privileged information to either teacher or student itself. However, this additional input induces a potential failure mode we dub **privilege illusion**: a pattern that conflates the transferable capability gap that students are meant to close, and the information asymmetry gap that can only be mimicked but never replicated. This issue is further amplified by the inherent non-uniformity of token-level supervision, where only a small subset of tokens carries pivotal capability-bearing signals. To this end, we propose **DOPD**, an advantage-aware dual distillation paradigm that dynamically routes token-level supervision between privileged teacher and privileged student policies based on their advantage gap and relative probabilities. Each token receives supervision of different strength, objective, and strategy from either teacher or student itself, which transfers credible capability while simultaneously receiving auxiliary signals. Extensive experiments on both large language model (LLM) and vision-language model (VLM) settings demonstrate that DOPD consistently outperforms Vanilla OPD and other counterparts. Further results on stability, robustness, continual learning, and out-of-distribution tasks validate its superiority.

1 INTRODUCTION

Distillation, as a powerful paradigm for transferring the capabilities of a high-performing teacher policy into a suboptimal student policy, typically relies on off-policy trajectories, which may expose the student to state distributions that are misaligned with its own evolving behavior (Jin et al., 2026; Sanh et al., 2019; Hsieh et al., 2023; Yu et al., 2026c; Yao et al., 2026). By contrast, recent OPD paradigms address this limitation by rolling out from the current student policy and using the teacher to provide token-level supervisory signals (Lab, 2025; Song & Zheng, 2026; Xiao et al., 2026; Xu et al., 2026a). This formulation not only mitigates distribution shift, but also delivers dense per-token teacher supervision via student-sampled on-policy trajectories, yielding higher distillation efficiency and superior performance.

Although OPD has emerged as an effective post-training paradigm, its achievable upper bound is fundamentally constrained by the quality of the supervision source (Li et al., 2026b; Fu et al., 2026; Ko et al., 2026). As demonstrated in Figure 1, for standard strong-to-weak distillation (Ko et al., 2025; Lab, 2025; Jang et al., 2026; Zhang et al., 2026; Ye et al., 2026; Yang et al., 2026b; Hou et al., 2026; Wu et al., 2026), the student is encouraged to imitate a stronger teacher; for self distillation (Shenfeld et al., 2026; Hübotter et al., 2026; Zhao et al., 2026; Penaloza et al., 2026; Stein et al., 2026), *i.e.*, self-as-teacher pattern, the model improves by regularizing itself under different contexts or conditions. In both cases, the effectiveness of OPD implicitly relies on the assumption: **the supervision signals should reflect a learnable capability beyond the current student policy**. However, this assumption might be fragile (Li et al., 2026b; Song & Zheng, 2026; Yang et al., 2026a), especially when privileged information is introduced.

Privileged information, such as verified reasoning hints for LLMs (Penaloza et al., 2026; Zhao et al., 2026; Hübotter et al., 2026; Yang et al., 2026a) or structured visual annotations for VLMs (Yuan et al., 2026; Liu et al., 2026), can indeed improve the prediction distribution of teacher policy and raise the apparent ceiling of distillation. Nevertheless, the theoretical gains afforded by privileged information training do not necessarily translate into transferable supervisory signals. Rather, they may stem from a hitherto uncharacterized failure mode, *i.e.*, privilege illusion: the ostensible performance gap between teacher and student in fact conflates two fundamentally distinct components.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

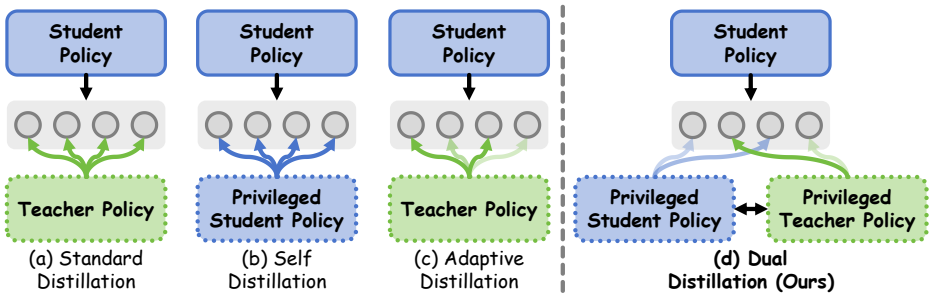


Figure 1: Comparison of existing **(a)** standard distillation, **(b)** self distillation, and **(c)** adaptive distillation paradigms with our proposed **(d)** dual distillation paradigm.

The first is the intrinsic teacher-student capability gap, which is expected to close through distillation; the second is a gap driven by information asymmetry, which arises from the access to privileged inputs that remain almost unlearnable. Indiscriminately distilling such a teacher distribution may therefore cause the student to fit privileged outcomes rather than acquire transferable ability. To summarize, **adding privileged inputs theoretically improves the ceiling, but the gains may stem from privilege illusion rather than capability optimization**, resulting in rapid entropy collapse, reduced exploration, and ultimately poor distillation effectiveness.

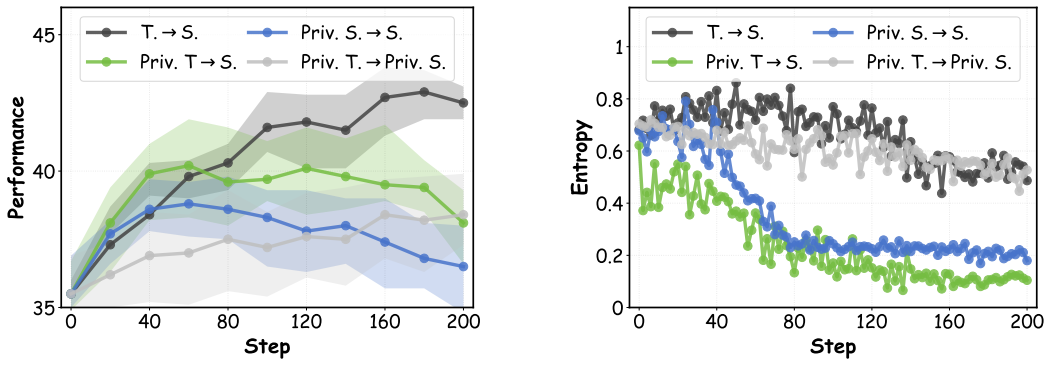
As the distillation signals are highly non-uniform across tokens, the concern of privileged illusion becomes more pronounced. For realistic trajectories, only a small subset of tokens may encode decisive branches, grounded evidence, critical preferences and other capacity-centric information (Xu et al., 2026b; Jin et al., 2026), while many others might provide low-value supervision, which might be privilege-dependent. However, the vanilla and most variants of OPD methods often optimize all tokens with the same supervision source and objective form, implicitly assuming that each token contributes equally to capability transfer (Lab, 2025; Li et al., 2026b; Song & Zheng, 2026). When incorporating privileged inputs, part of the teacher-student performance advantage originates from information gap rather than transferable capability. In this case, dense supervision might bias the student toward learning privilege-related shortcuts that are easier to fit than the underlying transferable capabilities, thereby amplifying the information-asymmetry component of the teacher-student gap. Thus, **indiscriminately and uniformly distilling all tokens from one monolithic policy might intensify the privilege illusion**.

Based on these insights, we propose an advantage-aware dual distillation paradigm, termed as **DOPD**, exploiting the complementary properties of teacher-based and self-based supervision under the privileged contexts to dynamically route token-level supervision between teacher and student policy according to the **privilege advantage gap** and their relative predicted probabilities. For tokens where the privileged teacher demonstrates a credible capability advantage, we apply stronger teacher distillation to transfer high-value privilege-conditioned capacity. As for tokens that are likely dominated by privileged information or less related to capacity, we instead rely on lighter supervision to preserve stability and encourage favorable exploration. In this way, dual distillation jointly adapts the supervision source, strength, and granularity, enabling more effective, stable, and adaptive distillation with less privilege illusion.

Extensive experiments demonstrate that our method achieves superior distillation performance across a wide range of scenarios, and exhibits excellent robustness, scalability, and generalization. Specifically, averaged across eight benchmarks, our method outperforms Vanilla OPD by **7.5** and **6.0 points** on LLM-based and VLM-based setups respectively, and sustains consistent improvements ranging from **6.2-10.6 points** across five model pairs of varying sizes. Furthermore, our method also delivers more favorable performance in continual learning, out-of-distribution evaluation, and training stability. Additional token and divergence analyses, sensitivity and ablation studies further corroborate the effectiveness and rationality of our approach.

2 Methodology

Vanilla OPD (Lab, 2025; Song & Zheng, 2026; Li et al., 2026b) trains a student policy on its own sampled trajectories using token-level supervision from a stronger teacher, mitigating the distri-



(a) Performance vs. Training Step

(b) Entropy vs. Training Step

Figure 2: Comparison of **(a)** performance and **(b)** entropy on variants with privileged information. Here, T./S., and Priv. denote teacher/student policy and with privileged information, respectively.

bution mismatch with denser signals. The supervision is typically defined through distributional divergences, including forward or reverse KL, and Jensen-Shannon divergence, and can be applied at different granularities, ranging from sampled-token and Top-K supervision to full-vocabulary. We defer detailed formulations and discussions of these objectives to Appendix B.1 and B.2.

2.1 Background

2.1.1 Privilege Illusion

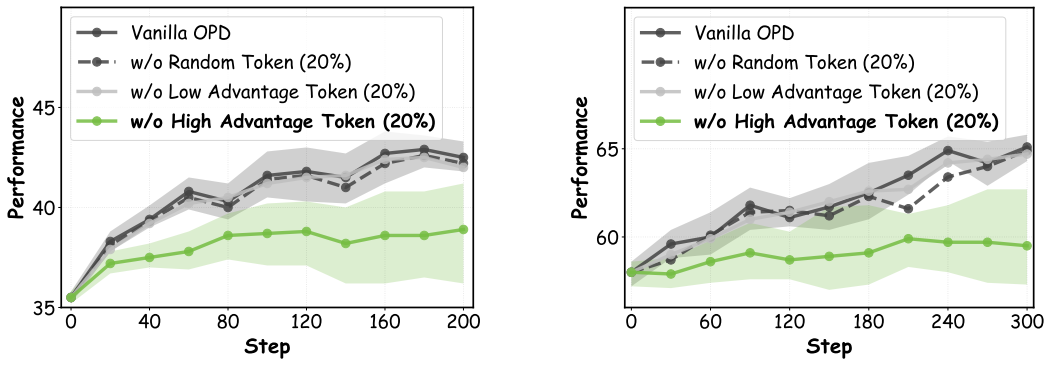
Existing OPD fundamentally relies on the assumption that a stronger teacher provides richer and more informative supervision (Lab, 2025; Ko et al., 2025). Thus, in many practical scenarios, an intuitive exploration is to equip teachers or student itself with privileged inputs (Song & Zheng, 2026). For instance, verified hints in reasoning-centric tasks (Zhao et al., 2026; Penaloza et al., 2026; Hübotter et al., 2026), or bounding boxes of objects in visual perception scenarios (Liu et al., 2026; Yuan et al., 2026). Here, we employ a moderate form of privileged information that delivers essential cues while refraining from directly disclosing detailed execution procedures and final answers. However, when augmented with privileged information, the prediction advantage may arise from information asymmetry rather than genuine inherent capability. Uncurated distilling such signals can encourage the student to imitate privileged outcomes instead of acquiring practical and transferable abilities, or triggers distillability due to irreparable teacher-student gap, leading to inferior and unstable distillation process, and unfavorable entropy collapse (Cui et al., 2025).

As illustrated in Figure 2, we compare the impact of privileged information inclusion on both performance and entropy trends. We evaluate three OPD variants, in which privileged information is granted to the teacher policy only, the student policy only, and both policies, respectively. We observe that introducing privileged information to either the teacher or the student separately delivers modest performance improvements over Vanilla OPD in the very early training phase, yet the information asymmetry between the two policies gives rise to late-stage performance degradation coupled with entropy collapse. When both policies are granted access to privileged information, the superficial advantage conferred by information asymmetry vanishes. Furthermore, uniform distillation across all tokens under this setting fails to enable the student to genuinely internalize the core competencies. Instead, the student merely passively adapts to the privileged information, ultimately yielding only marginal performance improvements less than Vanilla OPD.

In summary, the results reveal that straightforward incorporation of privileged information might create a failure phenomenon: privileged inputs may yield an ostensible advantage, however, such gains often stem from information asymmetry rather than from a genuine enhancement of capability.

2.1.2 Privilege Advantage Gap

As mentioned above, a key limitation of existing OPD methods is their inability to disentangle capability gaps from information gaps. Thus, we argue that, when both with privileged inputs,



(a) Qwen3-8B → Qwen3-1.7B (LiveBench) (b) Qwen3-VL-8B → Qwen3-VL-2B (MMStar)

Figure 3: Token ablations on random tokens, and tokens with high or low advantage gap.

the relative advantage between a teacher policy and a student policy offers a proxy for privilege-conditioned prediction gap. Consequently, a large advantage gap indicates capability discrepancy under controlled privileged conditions, whereas a small gap suggests that the advantage of teacher policy is primarily attributable to privileged information. This perspective motivates a privilege advantage-aware distillation paradigm that selectively transfers knowledge when the supervision signal reflects authentic competence rather than privilege illusion.

For a given input $\mathbf{x}$, the student policy samples an sequence from the conditional distribution. To conduct privilege advantage-aware distillation, we aim to investigate the distribution disparity between the teacher policy Π_T and student policy Π_S when both have access to privileged inputs, termed the privilege advantage gap $\mathcal{A}$. Then, using forward passes on the two policies respectively, and take the absolute value of their log-probability difference as privilege advantage gap:

$$\mathcal{A} = \left| \log \Pi_T \left(\mathbf{y}_n \mid \mathbf{x}, \mathbf{p}, \mathbf{y}_{<n} \right) - \log \Pi_S \left(\mathbf{y}_n \mid \mathbf{x}, \mathbf{p}, \mathbf{y}_{<n} \right) \right| = \left| \log \frac{\Pi_T \left(\mathbf{y}_n \mid \mathbf{x}, \mathbf{p}, \mathbf{y}_{<n} \right)}{\Pi_S \left(\mathbf{y}_n \mid \mathbf{x}, \mathbf{p}, \mathbf{y}_{<n} \right)} \right|, \quad (1)$$

where $\mathbf{y}_n$ denotes the current token to be evaluated by the two policies, and $\mathbf{p}$ denotes the privileged information provided as auxiliary contexts along with previous tokens. The quantity $\mathcal{A}$ captures the prediction discrepancy stemming from the performance gap between teacher and student policies under identical privileged conditions, which constitutes the idealized learning content.

To further verify the rationality of privilege advantage gap to separate capacity and information gap, we conduct comprehensive ablation studies and empirical analyses across both large LLMs and VLMs. Specifically, we construct three variants of the Vanilla OPD paradigm, each discarding particular tokens without distillation loss: (1) a reference baseline that randomly drops 20% of tokens; (2) a variant that prunes the 20% of tokens with the smallest advantage gap; (3) a variant that prunes the 20% of tokens with the largest advantage gap. As illustrated in Figure 3, ablating high-advantage tokens incurs substantial performance degradation and a marked reduction in distillation efficiency. At the 100th optimization step, removing high-advantage tokens achieves only approximately 50% of the performance gain obtained by Vanilla OPD. In contrast, pruning random or low-advantage tokens exerts negligible performance impact relative to the Vanilla OPD baseline. This performance disparity is even more pronounced in multimodal models, achieves only about 20% of the improvement achieved by Vanilla OPD. It is also noteworthy that despite underperforming relative to all counterparts, the variant with high-advantage tokens removed still yields tangible distillation gains by over 1.5 points, indicating that the remaining tokens, though less critical, remain indispensable.

2.1.3 TAKEAWAY

Based on these backgrounds, we summarize the following takeaway to support our method:

> Naively injecting privileged information can create a privilege illusion, where apparent gains arise from information asymmetry instead of transferable capability. The privilege advantage gap could highlight high-value tokens which are more critical for capacity-centric distillation.

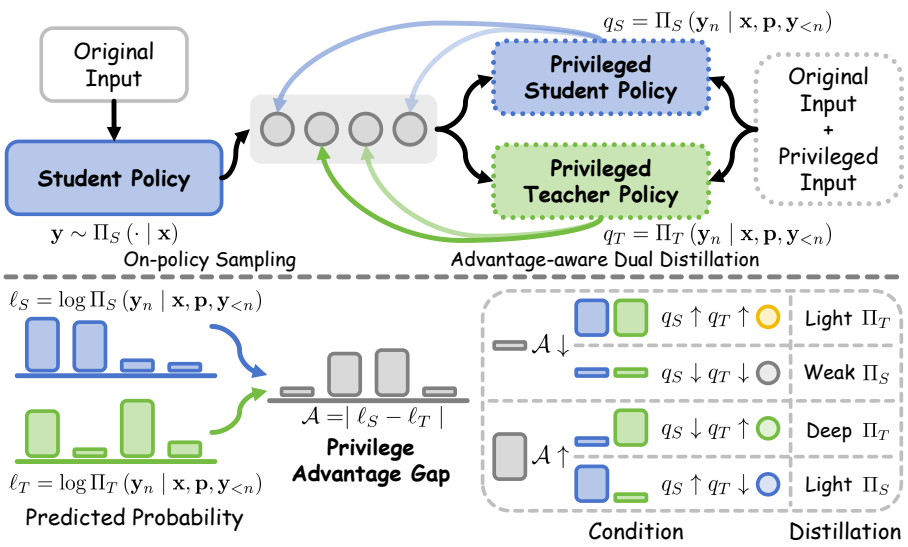


Figure 4: Overview of our proposed DOPD.

2.2 DOPD: Dual On-policy Distillation

As discussed above in Section 2.1, not all tokens should receive supervision of identical objective and strength or from the same source. When privileged information is introduced, the apparent superiority of teacher may originate either from privilege-conditioned capability discrepancy or from information asymmetry. Therefore, indiscriminately distilling the privileged teacher distribution may transfer shortcut-like privileged cues, while overly conservative self-teaching may fail to capture genuinely beneficial knowledge. To address this issue, we propose advantage-aware dual distillation, which dynamically selects both the supervision source and the distillation form according to the token-level privilege advantage gap.

Concretely, for each on-policy sampling trajectory $\mathbf{y}$, we perform additional privileged forward passes: one with the privileged student policy and the other with the privileged teacher policy. For the n-*th* token, we denote their token-level probabilities as: $q_S = \Pi_S(\mathbf{y}_n \mid \mathbf{x}, \mathbf{p}, \mathbf{y}_{<n})$ and $q_T = \Pi_T(\mathbf{y}_n \mid \mathbf{x}, \mathbf{p}, \mathbf{y}_{<n})$, while corresponding token-level log-probabilities as: $\ell_S = \log \Pi_S(\mathbf{y}_n \mid \mathbf{x}, \mathbf{p}, \mathbf{y}_{<n})$ and $\ell_T = \log \Pi_T(\mathbf{y}_n \mid \mathbf{x}, \mathbf{p}, \mathbf{y}_{<n})$. Here, the privileged student policy shares parameters with the deployed student policy, but receives the privileged input $\mathbf{p}$ during training, while the privileged teacher policy remains frozen. As we formally defined in Equation 1 of Section 2.1.2, we use the two privileged policies to calculate the privilege advantage gap $\mathcal{A}$. For a scored n-*th* token, we compare its $\mathcal{A}_n$, q_S, and q_T with their average $\bar{\mathcal{A}}$, $\bar{q_S}$, and $\bar{q_T}$, respectively. In practice, to ensure stability, we first discard the top 5% of outliers and perform normalization within the batch, then use them to calculate the average. Based on all these relationships, it yields four token regimes, each corresponding to a distinct learning strategy.

Low $\mathcal{A}$ with High q_S & q_T. When the two privileged policies have low advantage gap with both high predicted probability, *i.e.*, $\mathbb{I}^{\text{LH}} = (\mathcal{A}_n < \bar{\mathcal{A}}) \wedge (q_S + q_T \geq \bar{q_S} + \bar{q_T})$, the privileged teacher and privileged student make consistent and confident predictions. In this case, the bottleneck is mainly attributed to the absence of privileged information rather than an inherent capability gap. Thus, directly enforcing full teacher imitation is unnecessary and may over-transfer privileged shortcuts. We instead apply a light teacher distillation objective, using Top-K reverse KL to absorb useful privileged knowledge in a conservative manner:

$$\mathcal{L}^{LH} = \beta_l \, KL_{\text{reverse}}\left(\Pi_S(\cdot \mid \mathbf{x}, \mathbf{y}_{<n}) \,\|\, \Pi_T(\cdot \mid \mathbf{x}, \mathbf{p}, \mathbf{y}_{<n})\right). \tag{2}$$

where β_l denotes the intensity coefficient with light distillation.

Low $\mathcal{A}$ with Low q_S & q_T. When the two privileged policies have low advantage gap with both low predicted probability, *i.e.*, $\mathbb{I}^{\text{LL}} = (\mathcal{A}_n < \bar{\mathcal{A}}) \wedge (q_S + q_T < \bar{q_S} + \bar{q_T})$, both privileged policies assign low probability to the current token. Such tokens are likely to lie beyond the reliable competence region of both models, where aggressive teacher forcing may introduce noisy or even misleading

supervision. Therefore, we use the privileged student as a weak self-regularizing anchor, using Top-K reverse KL with a smaller coefficient, to stabilize training without forcing the student to imitate uncertain teacher predictions:

$$\mathcal{L}^{LL} = \beta_w \, KL_{\text{reverse}} \left(\Pi_S \left(\cdot \mid \mathbf{x}, \mathbf{y}_{<n} \right) \| \, sg \left[\Pi_S \left(\cdot \mid \mathbf{x}, \mathbf{p}, \mathbf{y}_{<n} \right) \right] \right), \tag{3}$$

where $sg[\cdot]$ denotes stop-gradient to avoid changing the gradient simultaneously to cause target drift, β_w denotes the intensity coefficient with weak distillation, and $\beta_w < \beta_l$. In this regime, the privileged student is not treated as a knowledge source, but as a parameter-shared consistency anchor that prevents policy drift.

High $\mathcal{A}$ with High q_T. When the two privileged policies have high advantage gap with high predicted probability of the teacher policy, *i.e.*, $\mathbb{I}^{\text{HT}} = \left(\mathcal{A}_n \geq \bar{\mathcal{A}} \right) \wedge \left(q_T \geq q_S \right)$, the privileged teacher exhibits a clear and confident advantage over the privileged student. Since both policies observe the same privileged information, a large privilege advantage gap suggests that, the teacher provides a potentially useful capability signal beyond what the student currently captures. Accordingly, these tokens contain critical transferable knowledge and should receive stronger supervision. We therefore perform full-vocabulary teacher distillation with unit weight, using JS divergence to balance support coverage and mode concentration:

$$\mathcal{L}^{HT} = JS \left(\Pi_S \left(\cdot \mid \mathbf{x}, \mathbf{y}_{<n} \right) \| \, \Pi_T \left(\cdot \mid \mathbf{x}, \mathbf{p}, \mathbf{y}_{<n} \right) \right). \tag{4}$$

Compared with Top-K strategy, full-vocabulary alignment provides denser distributional signals, enabling the student to acquire both dominant decisions and informative secondary preferences from the teacher.

High $\mathcal{A}$ with High q_S. When the two privileged policies have high advantage gap with high predicted probability of the student policy, *i.e.*, $\mathbb{I}^{\text{HS}} = \left(\mathcal{A}_n \geq \bar{\mathcal{A}} \right) \wedge \left(q_T < q_S \right)$, the privileged student assigns relative larger confidence while the privileged teacher does not provide a comparably reliable signal. In this regime, strongly constraining the student toward the teacher may suppress potentially valid exploratory behavior. We therefore adopt a light privileged-student distillation objective with Top-K reverse KL, which softly encourages consistency between the deployed student and its privileged counterpart while avoiding over-regularization:

$$\mathcal{L}^{HS} = \beta_l \, KL_{\text{reverse}} \left(\Pi_S \left(\cdot \mid \mathbf{x}, \mathbf{y}_{<n} \right) \| \, sg \left[\Pi_S \left(\cdot \mid \mathbf{x}, \mathbf{p}, \mathbf{y}_{<n} \right) \right] \right). \tag{5}$$

Total Objective. Finally, we combine the four token-wise objectives through indicator masks:

$$\mathcal{L}^{\text{DOPD}} = \mathbb{I}^{\text{LH}} \mathcal{L}^{\text{LH}} + \mathbb{I}^{\text{LL}} \mathcal{L}^{\text{LL}} + \mathbb{I}^{\text{HT}} \mathcal{L}^{\text{HT}} + \mathbb{I}^{\text{HS}} \mathcal{L}^{\text{HS}}, \tag{6}$$

where the masks are determined by the privilege advantage gap and relative probability comparisons described above, which exhaustively partitions the token space under the defined conditions. Thus, the overall optimization objective could be formulated as:

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[\mathbb{E}_{\mathbf{y} \sim \Pi_S} \left[\frac{1}{|\mathbf{y}|} \sum_{n=1}^{|\mathbf{y}|} \mathcal{L}_n^{DOPD} \left(\mathbf{y}_n ; \mathbf{x}, \mathbf{p}, \mathbf{y}_{<n} \right) \right] \right], \tag{7}$$

Through this adaptive routing mechanism, DOPD assigns strong full-vocabulary teacher supervision only to tokens where the privileged teacher demonstrates a credible capability advantage, applies light teacher distillation when the signal mainly reflects privileged information, relies on weak privileged-student regularization for uncertain regions, and preserves student exploration when the privileged student is already confident. Consequently, the proposed objective mitigates the entanglement between capability transfer and privileged-information imitation, yielding a more selective, stable, and generalizable OPD paradigm.

3 Experiments

We evaluate our proposed DOPD on eight text-based benchmarks and another eight for VLM-based tasks (detailed in Appendix C.1), which are conducted on Qwen3 and Qwen3-VL series with different sizes (detailed in Appendix C.2). For comprehensive comparisons, we include nine LLM-based OPD variants and four visual counterparts (detailed in Appendix C.3). In addition, the implementations are listed in Appendix C.4.

Table 1: Performance comparison of our proposed DOPD with counterparts on general, reasoning, and coding tasks. Δ indicates the performance gap between the student and teacher policies with gray cells, while blue for over 50% mitigation and green for complete gap removal by employing the OPD paradigms. The best and second best values are **bolded** and underlined, respectively.

Method	Type	General		Reasoning				Coding		Average
		C-Eval	LiveBench	MATH500	AIME25	ZebraLogic	AutoLogi	BFCLv3	LCBv5	
Teacher Policy	-	77.1	53.5	86.9	20.2	25.0	76.3	60.0	23.6	52.8
Student Policy		60.4	35.4	72.7	9.5	12.1	59.8	51.9	11.3	39.1
Δ		*+16.7*	*+18.1*	*+14.2*	*+10.7*	*+12.9*	*+16.5*	*+8.1*	*+12.3*	*+13.7*
Vanilla OPD	Standard	65.2	40.9	75.6	16.7	15.8	64.3	55.4	17.6	43.9
OPCD		66.1	41.6	75.3	19.5	17.0	65.2	54.7	15.8	44.3
ExOPD		68.3	44.7	76.7	18.5	19.9	68.0	57.2	22.6	47.0
Uni-OPD		66.5	42.3	77.5	20.0	22.3	67.2	56.1	20.8	46.6
SDFT	Self	63.4	38.7	73.8	15.0	15.4	62.6	53.2	12.1	41.8
OPSD		64.6	39.7	73.8	15.2	14.7	63.1	54.2	14.5	42.5
SDPO		65.4	38.8	74.0	16.4	15.1	62.9	52.3	17.7	42.8
EOPD	Adaptive	67.5	45.7	75.9	17.6	19.3	67.1	56.8	19.0	46.1
TIP		67.0	43.1	74.3	17.2	18.7	65.5	54.5	17.9	44.8
DOPD (Ours)	Dual	**71.3**	**49.8**	**81.5**	**23.3**	**26.9**	**71.0**	**60.2**	**27.1**	**51.4**

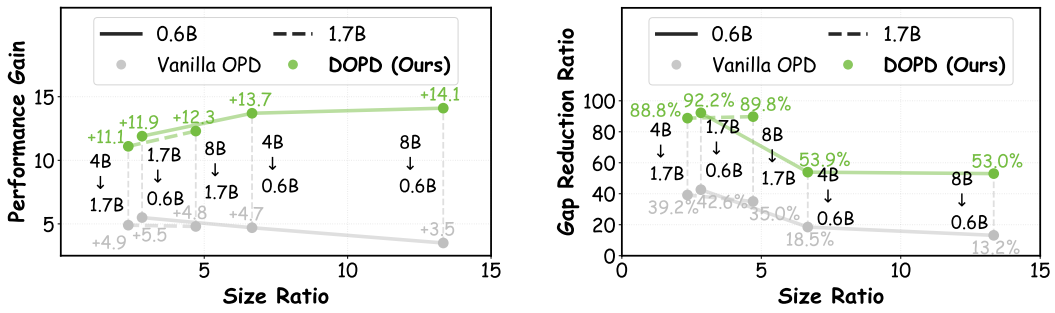
Table 2: Performance comparison of our proposed DOPD with counterparts on general, visual reasoning, and visual understanding tasks. $*$The codes of VA-OPD are not officially released, so we use the results of our reproduced version.

Method	General		Visual Reasoning			Visual Understanding			Average
	RealWorldQA	MMStar	MathVision	DynaMath	LogicVista	MMMU	MMMU-Pro	VSI-Bench	
Teacher Policy	71.3	70.7	53.8	67.6	55.0	69.6	55.8	59.7	62.9
Student Policy	63.6	58.4	32.0	53.8	35.5	53.2	36.4	53.6	48.3
Δ	*+7.7*	*+12.3*	*+21.8*	*+13.8*	*+19.5*	*+16.4*	*+19.4*	*+6.1*	*+14.6*
Vanilla OPD	64.7	61.8	37.1	56.2	40.2	58.0	46.7	54.1	52.4
Uni-OPD	65.0	65.3	43.0	58.2	42.5	59.1	47.0	53.7	54.2
Vision-OPD	66.2	66.4	38.0	57.6	43.1	64.9	52.3	56.1	55.6
VA-OPD*	67.0	66.2	38.7	57.7	43.1	66.1	**54.2**	57.5	56.3
DOPD (Ours)	**67.4**	**67.2**	**45.6**	**60.5**	**47.7**	**67.0**	53.9	**57.8**	**58.4**

3.1 Main Distillation Performance

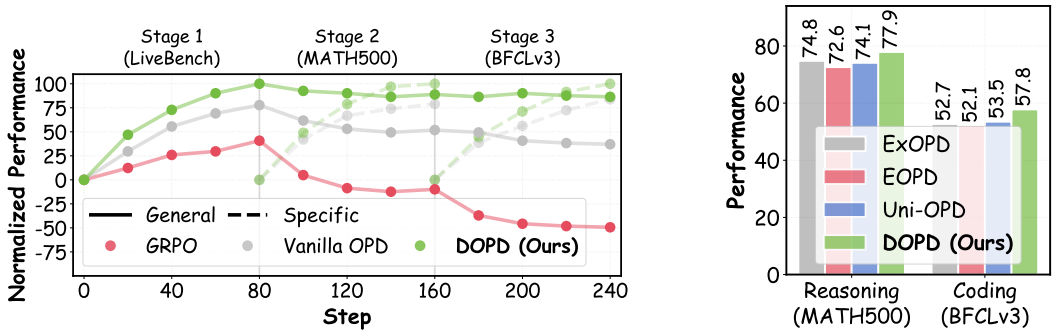
Across both LLM- and VLM-based OPD settings, as shown reported in Table 1 and 2, DOPD consistently and substantially narrows the teacher-student performance gap. On LLM benchmarks, DOPD improves the student policy by 12.3 points, recovering 89.8% of the original gap, and achieves the best performance across all eight benchmarks, outperforming the strongest adaptive distillation baselines, ExOPD (Yang et al., 2026b), Uni-OPD (Hou et al., 2026), and EOPD (Jin et al., 2026), by 4.4, 4.8, and 5.3 points on average, respectively. In contrast, self-distillation yields only modest gains, suggesting that relying solely on the student's own supervision is insufficient to effectively bridge the teacher-student gap. This advantage extends to VLM-based OPD, where DOPD improves the student policy by 10.1 points and recovers 69.2% of the gap, surpassing Vanilla OPD (Lab, 2025), Uni-OPD (Hou et al., 2026), Vision-OPD (Yuan et al., 2026), and VA-OPD (Liu et al., 2026) by 6.0, 4.2, 2.8, and 2.1 points, respectively. Notably, all methods exhibit larger gains on visual understanding than on reasoning and other visual tasks, potentially reflecting the visually grounded nature of the distillation paradigm, which primarily transfers accurate attention to and grounding in task-relevant visual evidence from the teacher.

These results demonstrate that advantage-aware dual distillation is more effective than either static teacher imitation, self-refinement, or single-sided adaptive weighting, indicating that DOPD transfers not only surface-level output preferences but also more essential ability from teacher policy. In addition, beyond text-only distillation, the proposed paradigm also provides robust and consistent gains for vision ability distillation.



(a) Performance Gain *vs.* Teacher-student Size Ratio (b) Gap Reduction *vs.* Teacher-student Size Ratio

Figure 5: Scalability comparison of proposed DOPD and Vanilla OPD on **(a)** performance gain and **(b)** teacher-student gap reduction ratio.



(a) Normalized Performance on Three-stage Continual Learning (b) Out-of-distribution Evaluation

Figure 6: Comparison of proposed DOPD and Vanilla OPD on **(a)** continual learning, where we conduct a three-stage continual learning with general, reasoning, and coding training sub-datasets sequentially; and **(b)** out-of-distribution tasks, where we optimize the student policy on coding or reasoning dataset, but evaluated on another out-of-domain benchmarks.

3.2 Robustness & Scalability

To assess the scalability of DOPD across different teacher-student configurations, we evaluate five model pairs with varying capacity gaps. As shown in Table 3, DOPD consistently outperforms Vanilla OPD (Lab, 2025) across all pairs, yielding gains of 11.1-14.1 points and achieving roughly two- to over three-fold larger improvements than Vanilla OPD. Notably, DOPD remains robust as the teacher-student size ratio increases, a regime known to exacerbate distribution mismatch and hinder effective policy transfer (Lab, 2025; Li et al., 2026b; Qin et al., 2026). For example, under the largest mismatch, Qwen3-8B $\rightarrow$ Qwen3-0.6B, Vanilla OPD improves the student by only 3.5 points, whereas DOPD achieves a 14.1-point gain and recovers 53.0% of the teacher-student gap, with a similar trend observed for Qwen3-4B $\rightarrow$ Qwen3-0.6B. As further illustrated in Figure 5a, Vanilla OPD becomes less effective as teacher capacity increases, while DOPD exhibits steadily increasing absolute gains, suggesting substantially better tolerance to capacity mismatch. Although the relative gap recovery inevitably decreases with larger size ratios due to the widening initial gap and the limited capacity of compact students (Figure 5b), DOPD markedly mitigates this degradation. Overall, these results demonstrate that DOPD provides a more scalable and reliable mechanism, particularly for transferring policies from substantially stronger teachers to compact students.

3.3 Continual Learning & Out-of-distribution Evaluation

Following prior work showing that OPD can better mitigate catastrophic forgetting than conventional post-training paradigms such as SFT and GRPO (Shenfeld et al., 2026; Hübotter et al., 2026; Shao et al., 2024), we conduct a three-stage continual learning experiment, sequentially introducing general, reasoning, and coding data. As shown in Figure 6a, OPD-based methods exhibit substantially stronger retention and sustained learning, with DOPD further amplifying this advantage:

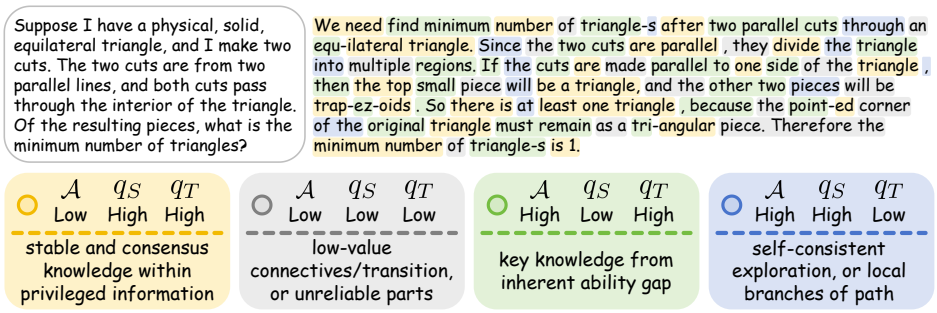


Figure 7: Token-level visualization of the four token types, where each token is colored based on their privilege advantage gap $\mathcal{A}$ and predicted probabilities of teacher q_T and student q_S policies.

performance consistently improves on each newly introduced domain while incurring only minor degradation on previously learned domains. This indicates that DOPD supports effective incremental capability accumulation rather than merely overwriting earlier knowledge.

We additionally assess cross-domain generalization by training on either reasoning or coding data and evaluating on the unseen complementary domain. As reported in Figure 6b, DOPD outperforms the strongest competing baselines, including ExOPD (Yang et al., 2026b), Uni-OPD (Hou et al., 2026), and EOPD (Jin et al., 2026), exceeding the second-best method by 3.1 and 4.3 points in the two transfer directions, respectively. Together, these results demonstrate that DOPD not only preserves previously acquired capabilities during sequential adaptation, but also learns policies that transfer more effectively beyond the training distribution.

3.4 Additional Analyses

We further conduct a series of analyses to understand the stability and key design choices of DOPD. First, as shown in Figure 9, DOPD exhibits consistently stable and efficient optimization throughout training, surpassing the final performance of representative standard, self-, and adaptive-distillation baselines at substantially earlier steps, while maintaining a well-behaved entropy trajectory without the collapse observed in self-distillation. Second, our privileged-information study shows that effective supervision depends not on exposing the final answer, but on providing capability-oriented guidance with an appropriate information gap: high-level step-wise hints are most effective for LLMs (Table 4), while localized object-level cues are preferable for VLMs (Table 5). As visualized in Figure 7 and Table 6, token-level analysis further reveals that different token categories play distinct roles during distillation; in particular, tokens that are likely under the teacher but unlikely under the student carry the most critical transferable knowledge, whereas indiscriminately distilling all tokens yields limited gains. As listed in Table 10, we also find that broader distribution alignment generally improves performance at increased computational cost, with hybrid configurations providing the strongest overall results in our setting. Finally, as demonstrated in Table 8 and 10, sensitivity and component ablations confirm that both privileged inputs, dual teacher-student supervision, and token-specific divergence designs are essential to the effectiveness of DOPD.

4 Conclusion

In this work, we revisit OPD under privileged contexts and identify fundamental limitations: the apparent superiority of a privileged teacher does not always correspond to transferable capability, but may instead arise from information asymmetry, and these supervision signals are not evenly distributed across tokens. Motivated by these observations, we propose DOPD, an advantage-aware dual on-policy distillation framework that adaptively routes token-level supervision between teacher-driven capability transfer and auxiliary self-optimization from the student. By leveraging the privilege advantage gap and relative token probabilities, DOPD selectively applies strong full-vocabulary teacher distillation to capability-bearing tokens, while imposing light or weak distillation on tokens without a capacity advantage gap. Extensive experiments across LLM and VLM settings demonstrate that DOPD consistently outperforms Vanilla OPD and strong competitive baselines, yielding superior distillation performance, robustness, continual-learning behavior, out-of-distribution generalization, and training stability.

5 AI USE STATEMENT

In this work, we used generative AI tools to improve the clarity, grammar, and presentation of our manuscript. All AI-assisted contents are carefully reviewed and verified by the authors. All scientific ideas, methodological designs, experimental results, and conclusions are developed and validated by the authors. We take full responsibility for the final content of our work.

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 (ICLR)*, volume 2024, pp. 21246–21263, 2024.

AIME. Aime problems and solutions, 2025. URL `https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions`.

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

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russ Webb. Distillation scaling laws. *arXiv preprint arXiv:2502.08606*, 2025.

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 (NeurIPS)*, 37:27056–27087, 2024.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, et al. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, et al. Uni-opd: Unifying on-policy distillation with a dual-perspective recipe. *arXiv preprint arXiv:2605.03677*, 2026.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, 2023.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:62991–63010, 2023.

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

Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 58791–58831, 2025.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. *arXiv preprint arXiv:2601.07155*, 2026.

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.

Minsang Kim and Seung Jun Baek. Explain in your own words: Improving reasoning via token-selective dual knowledge distillation. In *International Conference on Learning Representations (ICLR)*, 2026.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1317–1327, 2016.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. DistiLLM-2: A contrastive approach boosts the distillation of LLMs. In *International Conference on Machine Learning (ICML)*, 2025.

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.

Thinking Machines Lab. On-policy distillation, 2025. URL `https://thinkingmachines.ai/blog/on-policy-distillation`.

Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. Lavida: A large diffusion language model for multimodal understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:105101–105134, 2026a.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026b.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. Small models struggle to learn from strong reasoners. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25366–25394, 2025.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*, 2025.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, et al. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026.

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

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

Chuanyu Qin, Chenxu Yang, Qingyi Si, Naibin Gu, Dingyu Yao, Zheng Lin, Peng Fu, Nan Duan, and Jiaqi Wang. Near-future policy optimization. *arXiv preprint arXiv:2604.20733*, 2026.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

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

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

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Alex Stein, Furong Huang, and Tom Goldstein. Gates: Self-distillation under privileged context with consensus gating. *arXiv preprint arXiv:2602.20574*, 2026.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:95095–95169, 2024.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation. *arXiv preprint arXiv:2605.26844*, 2026.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, et al. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*, 4:2, 2024.

Yecheng Wu, Song Han, and Hai Cai. Lightning opd: Efficient post-training for large reasoning models with offline on-policy distillation. *arXiv preprint arXiv:2604.13010*, 2026.

xAI. Realworldqa: A benchmark for real-world spatial understanding, 2024. URL `https://huggingface.co/datasets/xai-org/RealworldQA`.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

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

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026a.

Wenda Xu, Rujun Han, Zifeng Wang, Long Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations (ICLR)*, 2025.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026b.

Fanjia Yan, Huanzhi Mao, Charlie Cheng-Jie Ji, Tianjun Zhang, Shishir G. Patil, Ion Stoica, and Joseph E.Gonzalez. Berkeley function calling leaderboard, 2024. URL `https://gorilla.cs.berkeley.edu/blogs/8_berkeley_function_calling_leaderboard.html`.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026a.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 10632–10643, 2025b.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026b.

Dingyu Yao, Junhao Zhou, Chenxu Yang, Chuanyu Qin, Haowen Hou, Zheming Liang, Congcong Wang, Yuhang Cao, Shenglong Ye, Shuai Xie, et al. Joyai-vl-interaction: Real-time vision-language interaction intelligence. *arXiv preprint arXiv:2606.14777*, 2026.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

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 (NeurIPS)*, 38:113222–113244, 2026a.

Xinlei Yu, Zhangquan Chen, Yongbo He, Tianyu Fu, Cheng Yang, Chengming Xu, Yue Ma, Xiaobin Hu, Zhe Cao, Jie Xu, et al. The latent space: Foundation, evolution, mechanism, ability, and outlook. *arXiv preprint arXiv:2604.02029*, 2026b.

Xinlei Yu, Chengming Xu, Guibin Zhang, Zhangquan Chen, Yudong Zhang, Yongbo He, Peng-Tao Jiang, Jiangning Zhang, Xiaobin Hu, and Shuicheng Yan. Vismem: Latent vision memory unlocks potential of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31544–31555, 2026c.

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.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL: Long Papers)*, pp. 15134–15186, 2025.

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler II, Qian Qian, Gregory D Lyng, Sanjit Singh Batra, and Robert E Tillman. Fast and effective on-policy distillation from reasoning prefixes. *arXiv preprint arXiv:2602.15260*, 2026.

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

Qin Zhu, Fei Huang, Runyu Peng, Keming Lu, Bowen Yu, Qinyuan Cheng, Xipeng Qiu, Xuanjing Huang, and Junyang Lin. Autologi: Automated generation of logic puzzles for evaluating reasoning abilities of large language models. *arXiv preprint arXiv:2502.16906*, 2025.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 48337–48383, 2025.

APPENDIX

A RELATED WORKS

A.1 TEACHER-STUDENT DISTILLABILITY

Teacher-student distillation has long been studied as a means of transferring capability from a stronger teacher to a weaker student model (Kim & Rush, 2016; Sanh et al., 2019; Yu et al., 2026b). In the era of large models, teacher–student distillability has become a more nuanced question than mere teacher imitation: recent work shows that teachers can transfer not only labels, but also rationales, trajectories, preferences, and broader behavioral patterns to smaller students (Hsieh et al., 2023; Gu et al., 2024; Agarwal et al., 2024). However, such transfer is not monotonic in teacher strength. Studies on the capacity gap suggest that an overly powerful teacher may provide signals that are difficult for a limited student to absorb (Busbridge et al., 2025; Li et al., 2025; Xu et al., 2025). Recently, some works also report this phenomenon on OPD settings, indicating that a more compatible initial distribution may be needed for better teacher-student distillation (Li et al., 2026b; Fu et al., 2026; Ko et al., 2026; Agarwal et al., 2024). Collectively, these studies suggest that effective distillation depends not only on a more powerful teacher model but also on the content and form of capacity being transferred.

A.2 ON-POLICY DISTILLATION

OPD has emerged as a compelling post-training paradigm that unifies the distributional consistency of on-policy learning with the dense supervision. As depicted in Figure 1, this field has evolved along three structured research directions: (a) standard distillation, *i.e.*, strong-to-weak paradigm, where a higher-capacity teacher model transfers knowledge to a weaker student via supervision on student-generated rollouts (Agarwal et al., 2024; Ko et al., 2025; Lab, 2025). Recent efforts are primarily structural modifications to the baseline to enable more stable, faster or more effective: Veto (Jang et al., 2026), Fast OPD (Zhang et al., 2026), OPCD (Ye et al., 2026), ExOPD (Yang et al., 2026b), Uni-OPD (Hou et al., 2026), Lightning OPD (Wu et al., 2026), Vision-OPD (Yuan et al., 2026), and VA-OPD (Liu et al., 2026). (b) self distillation, repurposing a single model as both teacher and student under different context conditions, including: SDFT (Shenfeld et al., 2026), SDPO (Hübotter et al., 2026), OPSD (Zhao et al., 2026), PI-Distill (Penaloza et al., 2026), RLSD (Yang et al., 2026a), and GATES (Stein et al., 2026). (c) adaptive distillation, which dynamically modulates supervision strategy based on student state, or other training signals: EOPD (Jin et al., 2026), TA-OPD (Wang et al., 2026), TIP (Xu et al., 2026b), REOPOLD (Ko et al., 2026), and TSD-KD (Kim & Baek, 2026).

Despite remarkable progress attained by such methods, they remain subject to fundamental limitations. In Vanilla OPD, student performance is subject to an inherent theoretical ceiling dictated by the performance of teacher policy (Song & Zheng, 2026; Li et al., 2026b; Lab, 2025). This constraint becomes particularly pronounced in challenging tasks, where the teacher itself exhibits subpar performance. While several lines of research have made preliminary attempts to leverage privileged information (Shenfeld et al., 2026; Yang et al., 2026a; Zhao et al., 2026; Hübotter et al., 2026; Penaloza et al., 2026; Stein et al., 2026; Yuan et al., 2026; Liu et al., 2026), these approaches generally operate under the implicit assumption that transferable capabilities can be enhanced via the direct integration of privileged information, and the supervision signals should be received uniform distillation mechanisms consistently from monolithic source. Critically, these methods fundamentally overlook the risk of privilege illusion, thus may fail to explicitly identify and distill genuine inherent capacity.

B METHODOLOGY

B.1 DIVERGENCE

As discussed in prior work (Hübotter et al., 2026; Jin et al., 2026; Li et al., 2026b), to learn a student from a teacher under the OPD framework, we first consider three common divergence-based objectives derived from Kullback-Leibler (KL) divergence: forward KL, reverse KL, and Jensen-

Shannon (JS) divergence. For notational simplicity, we omit the distinction between privileged and non-privileged observations and denote $\mathbf{t}$ as the current contexts of the teacher and student policies.

B.1.1 FORWARD KL DIVERGENCE

It encourages the student to cover the full support of the teacher distribution by penalizing actions that receive non-negligible probability under the teacher but are underestimated by the student. As a result, this objective promotes comprehensive imitation of the action preferences of teacher:

$$KL_{\text{forward}}\left(\Pi_T \,\|\, \Pi_S\right) = \mathbb{E}_{\mathbf{y}\sim\Pi_T(\cdot|\mathbf{t})}\left[\log \frac{\Pi_T\left(\mathbf{y} \mid \mathbf{t}\right)}{\Pi_S\left(\mathbf{y} \mid \mathbf{t}\right)}\right]. \tag{8}$$

B.1.2 REVERSE KL DIVERGENCE

It encourages the student to concentrate probability mass on actions strongly favored by the teacher, while assigning little emphasis to low-probability regions of the teacher distribution. Such mode-seeking behavior often leads to sharper student policies, but may also discard informative secondary modes encoded by the teacher:

$$KL_{\text{reverse}}\left(\Pi_S \,\|\, \Pi_T\right) = \mathbb{E}_{\mathbf{y}\sim\Pi_S(\cdot|\mathbf{t})}\left[\log \frac{\Pi_S\left(\mathbf{y} \mid \mathbf{t}\right)}{\Pi_T\left(\mathbf{y} \mid \mathbf{t}\right)}\right]. \tag{9}$$

B.1.3 JS DIVERGENCE

It introduces an intermediate average distribution, and calculate the KL divergence of teacher and student relative to this medium, without directional bias in forward or reverse directions. Their combination provides a more balanced optimization signal, thereby improving the stability of policy distillation:

$$JS = \frac{1}{2}KL\left(\Pi_T \,\|\, \Pi_M\right) + \frac{1}{2}KL\left(\Pi_S \,\|\, \Pi_M\right), \text{where} \quad \Pi_M = \frac{1}{2}\Pi_T + \frac{1}{2}\Pi_S. \tag{10}$$

B.2 OPD

As a promising post-training paradigm, OPD holds its core advantage in performing knowledge transfer with samples drawn from the target student policy to effectively mitigate performance bias caused by distribution shift, and provide richer supervision signals than conventional reinforcement learning paradigms (Lab, 2025; Song & Zheng, 2026; Li et al., 2026b). Specifically, given the student policy, for particular inputs $\mathbf{x}$, it samples the sequence of predicted trajectory $\mathbf{y} \sim \Pi_S\left(\cdot \mid \mathbf{x}\right)$. Then, the teacher policy, typically a stronger model, will offer token-level signals as optimization. Thus, the optimization objective of Vanilla OPD could be summarized as:

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbb{E}_{\mathbf{y}\sim\Pi_S}\left[\frac{1}{|\mathbf{y}|}\sum_{n=1}^{|\mathbf{y}|}\mathcal{L}_n\left(\mathbf{y}_n; \mathbf{t}_{<n}\right)\right]\right], \tag{11}$$

where $\mathbf{t}$ denotes the conditioning context, which comprises the original inputs, previously generated tokens, and auxiliary information if available, and $\mathcal{L}$ quantifies the token-level divergence between the teacher and student policies. Conventionally, this penalty term takes the form of divergence-based objectives, *e.g.*, widely adopted reverse KL, as well as alternative divergence variants or combinations thereof. Fundamentally, nearly all advancements in OPD center on minimizing the objective formalized in Equation 11, so as to yield a student model whose behavioral distribution aligns more closely with that of the teacher.

In addition, OPD approaches have different granularity of teacher supervision, ranging from coarse to fine: sampled-token, Top-K token, and full-vocabulary distillation. Sampled-token distillation confines its distillation objective exclusively to the predicted target token, while Top-K token distillation expands the scope of supervision to cover the k tokens with the highest predictive probabilities. By contrast, full-vocabulary distillation aligns the complete probability distribution across the entire vocabulary. The density of informative supervisory signals increases monotonically, which theoretically leads to higher efficiency, however, this gain comes at the cost of higher computational

overhead and potential risk of training instability, which stems from overfitting to the inherently noisy distributions of low-probability tokens (Zhao et al., 2026; Li et al., 2026b; Hübotter et al., 2026). Accordingly, the selection of distillation paradigm in practical deployment is typically tailored to specific downstream objectives and computational budgets.

C Experimental Settings

C.1 Benchmarks

To evaluate the effectiveness of our method, we employ eight benchmarks for LLM-based OPD, covering three core abilities: (1) general: C-Eval (Huang et al., 2023), and LiveBench (White et al., 2024); (2) reasoning: MATH500 (Hendrycks et al., 2021), AIME25 (AIME, 2025), ZebraLogic (Lin et al., 2025), and AutoLogi (Zhu et al., 2025); and (3) coding: BFCLv3 (Yan et al., 2024), and LCBv5 (Jain et al., 2025). For VLM-based OPD, we also include eight benchmarks on three aspects: (1) general: RealWorldQA (xAI, 2024), and MMStar (Chen et al., 2024); (2) visual reasoning: MathVision (Wang et al., 2024), DynaMath (Zou et al., 2025), and LogicVista (Xiao et al., 2024); and (3) visual understanding: MMMU (Yue et al., 2024), MMMU-Pro (Yue et al., 2025), and VSI-Bench (Yang et al., 2025b). All benchmarks are evaluated using their official metrics and evaluations to ensure fair and consistent comparison.

C.2 Models

We perform all the experiments on Qwen3 (Yang et al., 2025a) and Qwen3-VL (Bai et al., 2025) families of non-thinking versions as both teacher and student policies. Specifically, the main experiments and all analyses are conducted on Qwen3-8B to Qwen3-1.7B pair, and for VLM scenario is based on Qwen3-VL-8B to Qwen3-VL-2B pair. Besides, to verify the generalization ability of our method, we also add Qwen3-8B to Qwen3-0.6B, Qwen3-4B to Qwen3-0.6B, Qwen3-4B to Qwen3-1.7B, and Qwen3-1.7B to Qwen3-0.6B pairs.

For the training datasets of LLM-based OPD, we use the high-quality mixture dataset from RaR-Science-20K (Gunjal et al., 2025), DAPO-Math-17K (Yu et al., 2026a), and Skywork-OR1-Coding-14K (He et al., 2025), covering general, reasoning, and coding tasks. For VLM-based training datasets, we utilize ViRL39K (Li et al., 2026a) dataset, covering general, visual reasoning and visual understanding tasks. For the corresponding privileged input, we use GPT-5.4 (OpenAI, 2026) (2026-03-05) to generate step-wise decomposition hints and structured visual annotations respectively, where the generation prompts are provided in Figure 13. As illustrated in Figure 11, for LLM tasks, we utilize verified rationales as privileged information, with step-wise decomposition hints, but without direct execution trace or final answer. While as shown in Figure 12, for VLM tasks, privileged information denotes structured visual annotations, here we use query-related bounding boxes, with object labels and quadruple coordinates to provide explicit visual context. To guarantee the data quality, we use GPT-5.4 again to recheck the generated privileged contents, and directly discard relatively low-quality samples, eventually resulting in 32K and 25K high-quality training data for LLM and VLM, respectively.

C.3 Baselines

We compare our DOPD with other nine LLM-based counterparts, including three main paradigms of OPD as we discussed in Section A: (a) standard distillation: Vanilla OPD (Lab, 2025), OPCD (Ye et al., 2026), ExOPD (Yang et al., 2026b), and Uni-OPD (Hou et al., 2026); (b) self distillation: SDFT (Shenfeld et al., 2026), OPSD (Zhao et al., 2026), and SDPO (Hübotter et al., 2026); and (c) adaptive distillation: EOPD (Jin et al., 2026) and TIP (Xu et al., 2026b). For VLM-based methods, we benchmark DOPD with other four methods: Vanilla OPD (Lab, 2025), Uni-OPD (Hou et al., 2026), Vision-OPD (Yuan et al., 2026), and VA-OPD (Liu et al., 2026). For fair comparison, we rerun all the baselines on Qwen3/Qwen3-VL models.

Table 3: Generalization comparison of our proposed DOPD and Vanilla OPD based on five pairs of teacher-student models, including Qwen3-8B/4B/1.7B → Qwen3-0.6B, and Qwen3-8B/4B → Qwen3-1.7B.

Model Pair	Method	General		Reasoning				Coding		Average
		C-Eval	LiveBench	MATH500	AIME25	ZebraLogic	AutoLogi	BFCLv3	LCBv5	
Base Model	Qwen3-8B	77.1	53.5	86.9	20.2	25.0	76.3	60.0	23.6	52.8
	Qwen3-4B	72.2	48.3	84.6	18.9	35.0	75.8	57.4	21.3	51.6
	Qwen3-1.7B	60.4	35.4	72.7	9.5	12.1	59.8	51.9	11.3	39.1
	Qwen3-0.6B	42.0	21.8	55.2	2.0	4.1	37.2	44.0	3.5	26.2
Qwen3-8B	*Δ Performance Gap*	*+35.1*	*+31.7*	*+31.7*	*+18.2*	*+20.9*	*+39.1*	*+16.0*	*+20.1*	*+26.6*
↓	Vanilla OPD	44.8	24.5	59.3	4.6	5.9	46.5	46.1	6.1	29.7
Qwen3-0.6B	**DOPD (Ours)**	**56.7**	**35.5**	**70.9**	**17.0**	**19.6**	**54.3**	**55.5**	**12.7**	**40.3**
Qwen3-8B	*Δ Performance Gap*	*+16.7*	*+18.1*	*+14.2*	*+10.7*	*+12.9*	*+16.5*	*+8.1*	*+12.3*	*+13.7*
↓	Vanilla OPD	65.2	40.9	75.6	16.7	15.8	64.3	55.4	17.6	43.9
Qwen3-1.7B	**DOPD (Ours)**	**71.3**	**49.8**	**81.5**	**23.3**	**26.9**	**71.0**	**60.2**	**27.1**	**51.4**
Qwen3-4B	*Δ Performance Gap*	*+30.2*	*+26.5*	*+29.4*	*+16.9*	*+30.9*	*+38.6*	*+13.4*	*+17.8*	*+25.4*
↓	Vanilla OPD	44.8	24.0	60.9	7.4	7.9	47.4	46.5	8.6	30.9
Qwen3-0.6B	**DOPD (Ours)**	**54.0**	**33.8**	**72.7**	**16.3**	**25.1**	**51.6**	**52.2**	**13.7**	**39.9**
Qwen3-4B	*Δ Performance Gap*	*+11.8*	*+12.9*	*+11.9*	*+9.4*	*+22.9*	*+16.0*	*+5.5*	*+10.0*	*+12.5*
↓	Vanilla OPD	65.3	41.0	76.3	17.1	16.8	65.6	53.5	16.6	44.0
Qwen3-1.7B	**DOPD (Ours)**	**69.9**	**48.1**	**80.2**	**22.0**	**29.1**	**70.8**	**57.2**	**24.4**	**50.2**
Qwen3-1.7B	*Δ Performance Gap*	*+18.4*	*+13.6*	*+17.5*	*+7.5*	*+8.0*	*+22.6*	*+7.9*	*+7.8*	*+12.9*
↓	Vanilla OPD	47.3	29.5	60.1	7.8	9.9	43.3	47.1	8.2	31.7
Qwen3-0.6B	**DOPD (Ours)**	**55.6**	**35.6**	**66.7**	**14.6**	**16.4**	**49.4**	**53.0**	**13.7**	**38.1**

C.4 Implementations

All experiments are conducted on 8 NVIDIA H200 141GB GPUs. During distillation, the teacher policy is frozen for stability, while the student policy is optimized by AdamW optimizer and cosine scheduler with a learning rate of 5×10^{-6}. The batch sizes are set to 128 and 64 for LLM and VLM with 4 rollout samples, optimizing for a maximum of 200 and 300 steps, respectively. The K is set to 128 for Top-K distillation, and β_w and β_l are 0.3 and 0.6 to regulate the strength of distillation.

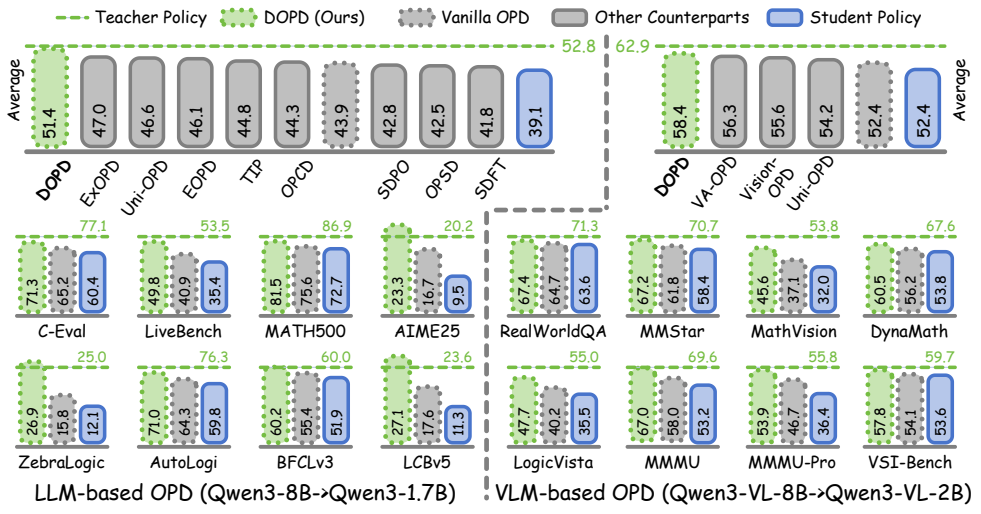
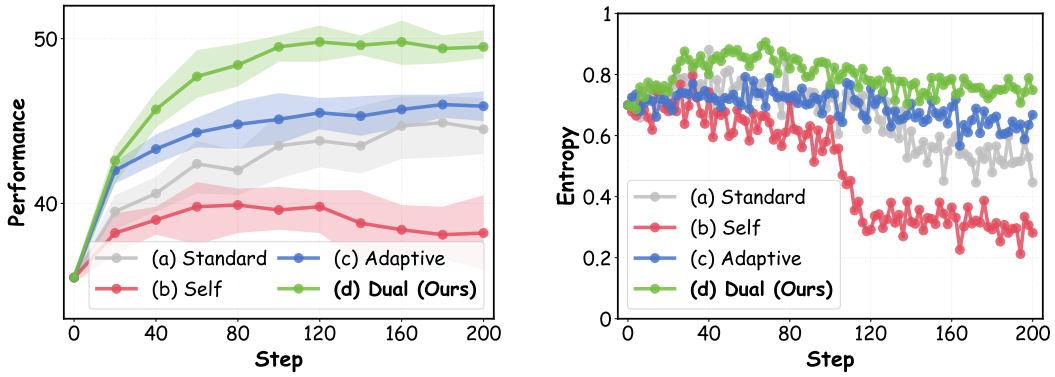


Figure 8: Performance comparison of our DOPD with competing approaches across eight benchmarks in terms of average across all benchmarks (upper bigger bars) and individual values of each benchmark (lower small bars).



(a) Performance *vs.* Training Step

(b) Entropy *vs.* Training Step

Figure 9: Training stability comparison of proposed DOPD and representative baselines, reporting the **(a)** performance and **(b)** entropy trends over training steps on LiveBench.

Table 4: Comparison of various LLM-based privileged information incorporation.

Privileged Input	C-Eval	LiveBench
Final Answer	59.5	36.7
Step-wise Hints with Execution	63.1	38.9
Step-wise Hints without Execution	**71.3**	**49.8**
Summarized Hints	65.8	43.6
No Privileged Input	63.0	39.4

Table 5: Comparison of various VLM-based privileged information incorporation.

Privileged Input	RealWorldQA	MMStar
Final Answer	64.6	63.2
Bounding Box with Caption	65.3	66.9
Bounding Box with Object Label	**67.4**	**67.2**
Caption	64.8	65.6
No Privileged Input	63.2	60.0

D Additional Experiments

D.1 Training Stability

To further assess training stability, we benchmark our method against the best-performing baselines from three distinct distillation paradigms: ExOPD (Yang et al., 2026b) for standard distillation, SDPO (Hübotter et al., 2026) for self-distillation, and EOPD (Jin et al., 2026) for adaptive distillation. As depicted in Figure 9a, our method consistently delivers stable and superior performance throughout the entire training process, coupled with higher distillation efficiency. Compared with the three competing paradigms, our method surpasses their step-200 performance as early as step-80. As shown in Figure 9b, our method maintains a healthy entropy trajectory: it rises modestly in the early training stage, followed by a gradual decline, and converges to a steady state after step-110. This pattern reflects that the model undergoes stable learning with well-calibrated exploration. Notably, we observe that the self-distillation paradigm encounters entropy collapse around step-95, alongside a subsequent drop in performance. This degradation is likely attributable to the insufficient and overly homogeneous supervision signals inherent to this paradigm, which render the learned distribution deficient in necessary exploration. Collectively, these results corroborate that our proposed method achieves superior performance gains in a stable and efficient manner throughout the distillation process.

D.2 Privileged Information Analysis

To evaluate the impacts of distinct privileged information injection strategies, we conduct comparative experiments to benchmark the performance of five different privileged information formulations: final answer, step-wise hints with detailed execution process, step-wise hints without execution, simplest summarized hints, and no privileged input for LLM-based distillation, and final answer, bounding box with descriptive caption, bounding box with object label, caption, and no privileged information for VLM-based task. As summarized in Table 4 and 5, directly providing ground-truth answers incurs the most severe information gap. The student model can only rigidly overfit to the given answers, which induces potential shortcut learning and performance degradation,

even underperforming the baseline without any privileged information. In contrast, providing only step-level high-level hints without detailed execution steps yields the largest LLM distillation gains of 8.3 and 10.4 points respectively. Meanwhile, providing bounding boxes paired with corresponding object labels proves to be the most suitable privileged information modality for VLM, bringing 4.2 and 7.2 points of improvement over the baseline. Notably, the efficacy of privileged information does not lie in the correctness of the final answers, but rather in its ability to deliver capability-oriented guidance to the student model, consistent with our previous discussion in Section 2.1.

D.3 Token Analysis

As detailed and analyzed in Section 2.1 and 2.2, we first compute the privilege advantage gap $\mathcal{A}$ and the predicted probabilities $q_{T/S}$ of both the teacher and student policies for each token, based on which we categorize each token into distinct classes. To intuitively characterize the functional roles of different token types during distillation, Figure 7 visualizes the distribution of token categories within a real trajectory. Among low-gap tokens, those with both high probabilities typically correspond to stable and consensus knowledge within privileged information, whereas tokens with both low probabilities are mostly connectives, transitions or unreliable segments with little valid information. Among high-gap tokens, tokens with high teacher probability but low student probability generally represent key knowledge arising from the inherent privilege-conditioned ability gap, while tokens with high student probability but low teacher probability likely reflect self-consistent or local branches of exploration. This token distribution pattern aligns well with our proposed token-level differentiated distillation strategy, enabling targeted and efficient distillation for tokens with distinct functional roles.

To further quantitatively dissect the contributions of individual token types and the adaptive advantage-aware dual distillation mechanism to our proposed method, we conduct token-level ablation analysis. Each variant performs distillation with signals from only one or combinations of token types, utilizing JS divergence on Top-K tokens. The setting without adaptive distillation corresponds to a baseline where all tokens receive identical distillation weights and strategies, with no token-wise differentiation. As listed in Table 6, using exclusively tokens with high teacher probability and low student probability already outperforms the equal-distillation setup using all four token types (equivalent to Vanilla OPD) by 4.6 points. However, naively adding the other three token types under an equal distillation scheme yields only marginal performance gains, and may even cause performance degradation. In contrast, equipping the framework with the adaptive distillation mechanism allows for adjustment of token-level distillation intensity, supervision granularity, and distillation content. These designed patterns render the distillation process more efficient and stable, delivering an overall improvement of over 8 points than equal distillation, when all four token types are leveraged.

D.4 Divergence Analysis

To further investigate the impacts of different divergence objectives (forward KL, reverse KL, and JS divergence) and strategies (sampled token, Top-K tokens, and full vocabulary), all introduced in Section 2.2, we conduct additional comparative experiments. To isolate the effects of other factors, we apply equal distillation across all tokens. Table 7 summarizes how these design choices shape final distillation performance and efficiency. Specifically, as the alignment scope expands from sampled to Top-K tokens and further to full vocabulary, performance improves progressively, yet inevitably incurs higher computational memory overhead. Furthermore, in contrast to findings reported in some prior works (Zhao et al., 2026; Jin et al., 2026), JS divergence delivers relatively superior performance than forward or reverse KL methods under our settings. Collectively, these results illustrate the inherent trade-off across different divergence configurations, providing empirical justification for our differentiated distillation paradigms.

D.5 Sensitivity & Ablation Studies

As illustrated in Figure 10, we conduct an analysis focusing on the distillation intensity assigned to different token categories. We observe that setting $\beta_w = 0.3$ and $\beta_l = 0.6$ strikes a favorable trade-off across token-wise distillation strengths: it amplifies the contribution of critical tokens while

Table 6: Effectiveness of individual or combinations of four tokens, and adaptive mechanism on LiveBench.

○	○	○	○	Adaptive	Step-40	Step-80	Step-160
✓	✗	✗	✗	✗	41.0	43.7	45.3
✓	✓	✗	✗	✗	39.9	42.9	45.0
✓	✗	✓	✗	✗	38.6	41.0	42.7
✓	✗	✗	✓	✗	39.2	42.7	44.5
✓	✓	✓	✓	✗	38.4	40.9	41.3
✓	✓	✗	✗	✓	41.0	45.4	47.9
✓	✗	✓	✗	✓	39.7	42.3	44.8
✓	✗	✗	✓	✓	40.1	44.9	47.6
✓	✓	✓	✓	✓	**45.7**	**48.4**	**49.8**

Table 7: Impact of different divergence objectives and strategies on LiveBench.

Objective	Strategy	Step-40	Step-80	Step-160
Forward KL	Sampled Token	37.2	39.8	41.1
	Top-K Tokens	38.5	40.2	41.3
	Full Vocabulary	38.0	40.1	41.5
Reverse KL	Sampled Token	37.0	38.4	40.6
	Top-K Tokens	37.8	39.0	40.8
	Full Vocabulary	38.5	40.6	41.9
JS Divergence	Sampled Token	37.3	39.2	41.0
	Top-K Tokens	38.4	40.9	41.3
	Full Vocabulary	39.2	41.6	42.5

Table 8: Ablation study on our DOPD, covering the main designs of advantage-aware dual distillation.

Variant	C-Eval	LiveBench
w/o Privileged Input	63.6	38.3
w/o Distillation from Student Policy	70.4	47.9
w/o Distillation from Teacher Policy	65.9	41.2
w/o Advantage-aware Distillation	67.6	41.3
w/o Adaptive Divergence Objectives	70.0	46.7
w/o Adaptive Divergence Strategies	70.8	46.1
DOPD (Ours)	**71.3**	**49.8**

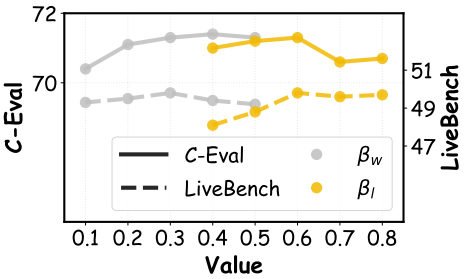


Figure 10: Sensitivity study on the intensity coefficient of weak β_w and light β_l distillation.

preserving the auxiliary role of other tokens in stabilizing and providing additional optimization signals.

Furthermore, to further disentangle the contributions of individual design components in our framework, we conduct ablation studies on two core elements: the sources of distillation signals and divergence-based designs. As presented in Table 8, privileged input is indispensable to our paradigm, as it directly underpins the advantage-aware calculation of our approach. Signals derived from the teacher policy serve as the primary driver of performance gains, while the student policy also fulfills an irreplaceable role throughout the distillation process. In addition, our token-wise divergence design tailored for distinct token categories is empirically validated to be effective.

E Details of Privileged Input

F Limitations and Future Directions

Notwithstanding the efficacy of our proposed DOPD, we acknowledge that several minor limitations remain. First, our method hinges on the availability and quality of privileged information, the construction of which incurs additional costs for annotation, generation, and filtering processes. Second, it introduces extra computational overhead relative to Vanilla OPD, requiring one additional forward pass of the student model. Third, while the current routing strategy is intuitive, and empirically stable, it still relies on heuristic mechanisms.

Future research may further advance DOPD along directions: developing more reliable and cost-effective mechanisms for obtaining privileged information or discovering alternative strategy to detect available advantage gap, with more principled or learnable distillation routing. More broadly, the paradigm of dynamic distillation from both teacher and student offers a useful lens for selective capacity transfer beyond LLMs and VLMs, inviting future work on more interpretable, efficient, and trustworthy distillation paradigms.

Original Input:
Suppose I have a physical, solid square pyramid. The bottom square has vertices A, B, C, D, and the final vertex is E. Then I make a cut through the plane defined by ACE. There are now two pieces. What are the pieces? Are they tetrahedra, squares, square pyramids, circles, triangular prisms, rectangular prisms, or are they both different shapes?

Privileged Input:
1. Identify the plane determined by the two opposite base vertices and the apex.
2. Observe how this plane intersects the square base along a diagonal.
3. Use that diagonal to partition the base into two congruent triangular regions.
4. Extend each triangular base region to the common apex to determine the corresponding three-dimensional subsolid.
5. Compare each resulting piece by its vertices, edges, and triangular faces.

Case 1 (LLM-based)

Original Input:
Factor the following quadratic: $5x^2+90x-1040$. Please put your final answer in a $\boxed{}$

Privileged Input:
1. Check whether the quadratic has a common numerical factor, if so, simplify quadratic.
2. Use the coefficient structure to decide which appropriate factorization strategy.
3. Identify the needed pairwise relationship between two numbers for the middle-term split.
4. Indicate that the remaining expression can be factored into two linear binomials.

Case 2 (LLM-based)

Original Input:
Consider all words constituted by eight letters from $\\{C ,H,M, O\\}$. We arrange the words in an alphabet sequence.\nPrecisely, the first word is $CCCCCCCC$, the second one is $CCCCCCCH$, the third is $CCCCCCCM$, the fourth one is $CCCCCCCO$, ...,$ and the last word is $OOOOOOOO$.\na) Determine the 2017th word of the sequence?

Privileged Input:
1. Recognize that the alphabetic ordering induces a four-symbol positional system.
2. Assign each letter an ordered digit according to this alphabet.
3. Convert the requested ordinal position to a zero-based rank before processing.
4. Express this rank as an eight-place base-four representation, preserving leading positions.
5. translate each base-four digit back to its corresponding letter.

Case 3 (LLM-based)

Figure 11: Demonstrations of LLM-based privileged input.

Original Input:
Which is the main topic of the image:
A: A woman surfing,
B: A man skating,
C: A man surfing,
D: A woman skiting.

Privileged Input:

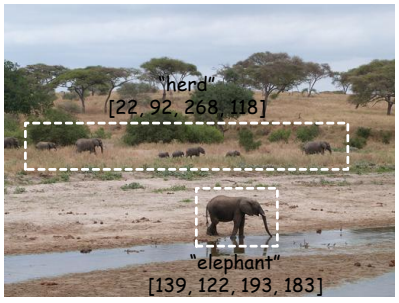


Case 1 (VLM-based)

Original Input:

Which one is the correct caption of this image?
A: A bunch of cars sitting still in the middle of a street,
B: Two giraffes near a tree in the wild,
C: Small personal bathroom with a tiny entrance door,
D: An elephant drinking water while the rest of the herd is walking in dry grass.

Privileged Input:



Case 2 (VLM-based)

Original Input:
What color are the flowers in the room?
A: White,
B: Yellow,
C: Red,
D: Pink

Privileged Input:



Case 3 (VLM-based)

Figure 12: Demonstrations of VLM-based privileged input.

```
Given a question, and corresponding ground-truth label.

Query: {Query}
Label: {Label}

Add necessary step-wise decomposition hints that support the answer.

Rules:
1. Please NOT provide the final answer, execution result, calculation result, or conclusion.
2. Please keep each hint concise and avoid overly detailed or case-specific explanations.
3. Please number the hints using Arabic numerals.

Output format:
1. ...
2. ...
3. ...
```

Prompt of Privileged Input Generation (LLM-based)

```
Given an image, a question, and corresponding ground-truth label.

Image: {Image}
Query: {Query}
Label: {Label}

Add necessary query-related bounding boxes that support the answer.

Rules:
1. Please add multiple boxes if necessary.

2. Please generate both the object label and quadruple coordinates.

3. Please output only valid JSON format without any other redundant content.

Output format:
[
  {"label": "object", "bbox": [x1, y1, x2, y2]}
]
```

Prompt of Privileged Input Generation (VLM-based)

Figure 13: Prompts of Privileged Input Generation.