FOCUSOPD: STRUCTURE-GUIDED ON-POLICY DISTILLATION FOR DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-teacher on-policy distillation enables rectified-flow models with diffusion-transformer architectures to combine task-specific capabilities within a single student. However, matching model outputs does not guarantee that their internal representations are aligned, and it remains unclear where and when additional feature supervision is most useful. We analyze the teacher–student feature mismatch across Transformer blocks and denoising timesteps, and find that the mismatch is concentrated in deeper Transformer blocks during early denoising and persists after OPD distillation, indicating ineffective distillation. Motivated by this observation, we introduce FocusOPD, which augments the base OPD objective with token-wise directional feature alignment restricted to deep layers and early denoising steps. Across two rectified-flow backbones and multiple OPD objectives, FocusOPD improves capability transfer in compositional generation, text rendering, and preference-based evaluation. On the primary benchmark, it improves over DiffusionOPD from 0.953 to 0.967 on GenEval, from 23.797 to 23.944 on PickScore, and from 6.052 to 6.199 on HPSv2.1. Ablations show that selective deep and early supervision is more effective and efficient than alignment over the full network and trajectory.

1 INTRODUCTION

On-policy distillation (OPD) transfers teacher behavior through feedback on student-generated trajectories (Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025). For diffusion models, multi-teacher OPD integrates specialized capabilities, including compositional generation, text rendering, and alignment with human preferences, into a single student (Ma et al., 2026; Li et al., 2026; Fang et al., 2026). Recent methods select denoising states for teacher feedback and align representation changes across network blocks (Zhou et al., 2026; Wei et al., 2026). Yet these developments leave open how the value of supervision varies across blocks and denoising stages. We therefore ask: *what patterns characterize the utility of additional feature supervision across network depth and denoising steps?*

To investigate this question, we begin by examining teacher–student feature mismatch in SD3.5-M Esser et al. (2024) and Z-Image (Z-Image Team et al., 2025). Evaluating teacher and student on the same noisy inputs, we find that feature mismatch is generally larger in deeper blocks and during early, high-noise denoising. Its relative organization remains qualitatively similar before and after output-level OPD, even though its magnitude may change. We call this a *persistent depth–temporal structure*. The marginal trends locate larger discrepancies; the joint profiles reveal that their organization persists through training. Neither observation alone explains the underlying mechanism.

We examine two mechanisms consistent with this structure. Two lemmas characterize how local discrepancies propagate across network blocks and how weaker observations preserve prior-dependent differences in a Gaussian denoising model. Building on this account, we derive a local criterion for selective feature supervision: under local smoothness and gradient-compatibility conditions, allocating a fixed auxiliary weight to sufficiently larger discrepancies can yield greater output-loss reduction for small auxiliary updates. Two lightweight diagnostics support the proposed mechanisms: upstream differences dominate mismatch at selected deep blocks, and noise-dependent scaling improves held-out prediction over a timestep control.

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.

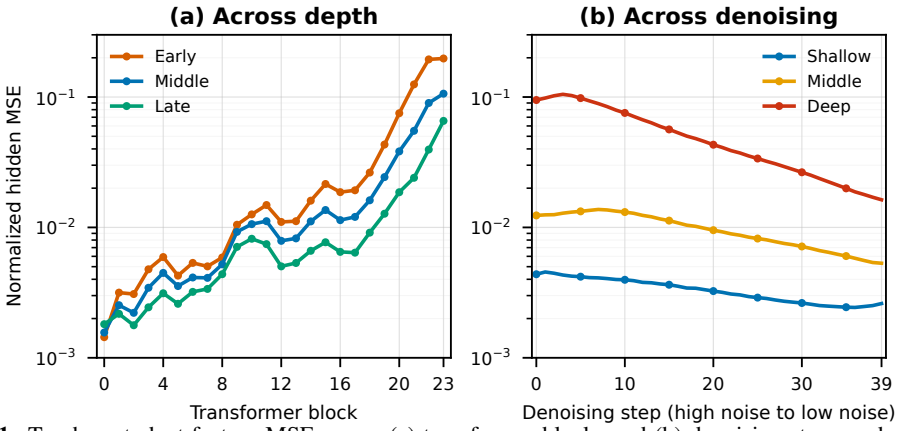


Figure 1: Teacher–student feature MSE across (a) transformer blocks and (b) denoising steps, ordered from high to low noise. Different curves indicates different groups of (a)denoising steps (b)blocks.

Motivated by this analysis, we introduce FocusOPD, a straightforward way to use the persistent depth–temporal structure in multi-teacher OPD (MOPD). It augments the base output objective with tokenwise cosine alignment of conditional image features in deep layers during early denoising. Output and feature queries are specified independently, and inference requires no additional components. On SD3.5-M with DiffusionOPD, it improves GenEval from 0.953 to 0.967, OCR accuracy from 0.958 to 0.960, and PickScore from 23.797 to 23.944, and ImageReward from 1.471 to 1.517. The ablations show that gains depend on supervision placement: Last-12/Front-40% reaches HPSv2.1 6.214 with a 432 MiB feature cache, whereas full-trajectory supervision yieds a GenEval score of 0.951, and supervising all blocks raises it only to 0.965 while doubling the cache. Our main contributions are:

- **Observation** (Section 1). We identify persistent structure of teacher-student mismatch across depth and denoising time in the MOPD of SD3.5-M and Z-Image.
- **Theory & Evidence** (Sections 3–4). We characterize two mismatch mechanisms, support them with targeted diagnostics, and derive a conditional local criterion for selective supervision.
- **Method** (Section 5). We introduce FocusOPD, aligning deep features in early denoising steps while retaining the base OPD protocol.
- **Experiments & Insights** (Section 6). We improve capability transfer across composition, text rendering, and preference metrics. The ablations show that selective supervision of deep layers during early denoising is effective, while broader temporal or depth coverage adds cost without consistent gains.

2 A Persistent Structure of Teacher–Student Mismatch

Output-level OPD transfers teacher behavior by matching denoising predictions. To understand the internal discrepancies that remain under this supervision, we begin by examining where teacher–student representations differ across network depth and denoising time, and whether their mismatch structure changes with training or varies on different models.

On SD3.5-M, we measure the MSE between corresponding conditional image-token features, evaluating teacher and student on the same noisy latent, timestep, and prompt. Figure 1 shows that mismatch is larger in deeper transformer blocks and during early, high-noise denoising steps.

These trends identify where discrepancies are larger, but do not explain how they arise. To further characterize this pattern, we examine its joint distribution across blocks and denoising steps before and after output-level OPD and extend the experiments with Z-Image.

Figure 2 reveals a consistent pattern across both backbones: larger discrepancies are concentrated in deeper blocks and earlier denoising steps. This organization persists after training, with only

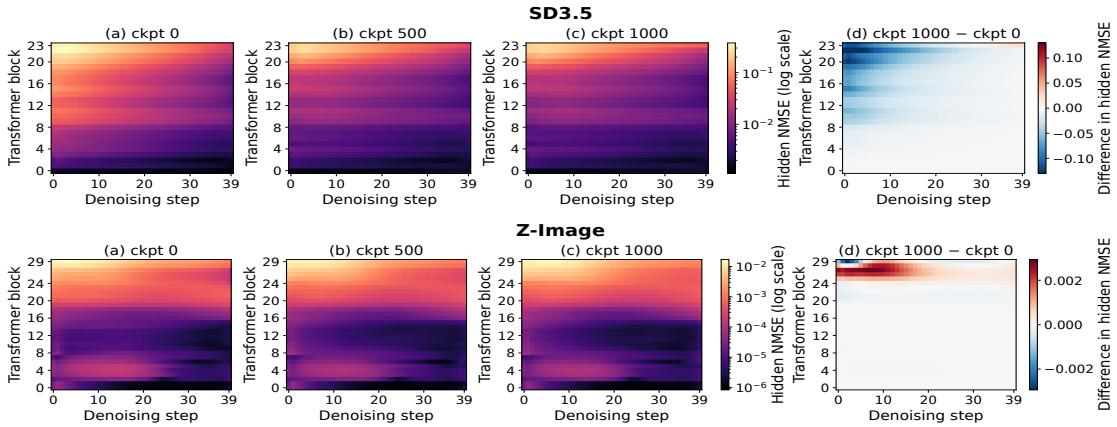


Figure 2: Hidden-NMSE heatmaps. Columns show checkpoints 0, 500, 1000, and the checkpoint-1000 minus checkpoint-0 difference.

small differences between training steps 0 and 1000, as shown in the rightmost column. We refer to this recurring pattern as a *persistent depth–temporal structure*. Persistence describes the relative distribution of mismatch, rather than an unchanged absolute MSE. This raises the question: *What gives rise to this structure, and why does it persist after output-level distillation?* Section 3 develops a mechanistic account of these observations.

3 From Mismatch Mechanisms to Selective Supervision

We first examine two mechanisms consistent with the observed structure, then analyze when mismatch can guide feature supervision. All proofs are in Appendix A.

Notation. Fix a student S and teacher T at the same noisy latent x_s and text condition c, with s indexing denoising from high to low noise. For $M \in \{S, T\}$, let $z^M_{\ell,s}$ be the complete input to block ℓ, including image/text tokens and side inputs, and $F^M_{\ell,s}$ its block map. Thus $z^M_{\ell+1,s} = F^M_{\ell,s}(z^M_{\ell,s})$; write $d_{\ell,s} = z^S_{\ell,s} - z^T_{\ell,s}$.

Lemma 3.1 (Depth-wise discrepancy transport)**.** *For continuously differentiable student blocks, define $b_{j,s} = F^S_{j,s}(z^T_{j,s}) - F^T_{j,s}(z^T_{j,s})$. Let $A_{j,s}$ be the student Jacobian averaged along the segment joining its teacher and student inputs, and set $P_{\ell\leftarrow j,s} = A_{\ell-1,s}\cdots A_{j,s}$, $P_{\ell\leftarrow\ell,s} = I$. Then*

$$d_{\ell,s} = P_{\ell\leftarrow 0,s} d_{0,s} + \sum_{j=0}^{\ell-1} P_{\ell\leftarrow j+1,s} b_{j,s}. \tag{1}$$

Deep mismatch thus combines local discrepancies with differences inherited from upstream states; their interaction determines accumulation.

Lemma 3.2 (Noise-dependent prior sensitivity)**.** *Let $y \mid c, M \sim \mathcal{N}(\mu_M(c), \tau^2 I)$ be a clean-latent prior with mean $\mu_M(c)$ and shared variance $\tau^2 > 0$. Observe $x_s = a_s y + \sigma_s \varepsilon$, with signal coefficient a_s, noise scale $\sigma_s > 0$, and $\varepsilon \sim \mathcal{N}(0, I)$ independent of y. The posterior means $\widehat{y}_M = \mathbb{E}_M[y \mid x_s, c]$ satisfy*

$$\widehat{y}_S - \widehat{y}_T = \rho_s\big(\mu_S(c) - \mu_T(c)\big), \qquad \rho_s = \frac{\sigma_s^2}{a_s^2\tau^2 + \sigma_s^2}. \tag{2}$$

The priors encode model-specific expectations under the same text. Lower $\mathrm{SNR}_s = a_s^2/\sigma_s^2$ preserves more of this prior difference. Together, Lemmas 3.1 and 3.2 motivate a structured view of mismatch across depth and denoising time. Building on this account, Proposition 3.3 establishes a conditional implication for feature supervision: under local smoothness and gradient-compatibility conditions, sufficiently larger mismatch can yield a greater single-step reduction in output loss at a matched supervision budget.

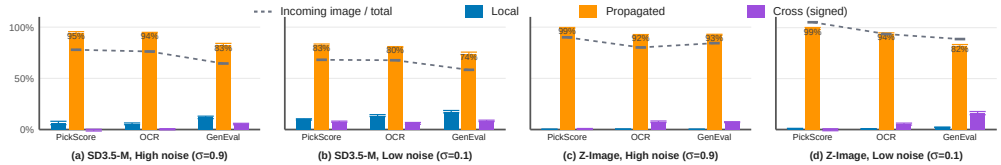


Figure 3: Deep-block decomposition across three teachers. Terms are normalized by total mean discrepancy energy. Block indices are 18 for SD3.5-M and 22 for Z-Image. Error bars: 95% prompt-bootstrap intervals.

Local supervision utility. Fix a replay batch and teacher targets for one update comparison. For position $i = (\ell, s)$, let ϕ_i^M stack the image features, with averaging factors and optional token normalization included. For student parameters θ, define $e_i(\theta) = \phi_i^S(\theta) - \phi_i^T$, $D_i = \|e_i\|_2^2$, and $\ell_i = D_i/2$. Let $g = \nabla_\theta \mathcal{L}_{\text{out}}$ be the output-loss gradient and $g_i = \nabla_\theta \ell_i$ the feature gradient, all evaluated at θ. A supervision distribution $p_i \geq 0$, $\sum_i p_i = 1$, defines $g_p = \sum_i p_i g_i$ and mean mismatch $S_p = \sum_i p_i D_i$. For a fixed learning rate $\eta > 0$ and auxiliary weight $\lambda > 0$, compare

$$\theta_{\text{out}}^+ = \theta - \eta g, \qquad \theta_p^+ = \theta_{\text{out}}^+ - \delta g_p, \qquad \delta = \eta\lambda. \tag{3}$$

Write $g_{\text{out}}^+ = \nabla \mathcal{L}_{\text{out}}(\theta_{\text{out}}^+)$ and define the correction efficiency $\alpha_i = \langle g_{\text{out}}^+, g_i \rangle / D_i$ for $D_i > 0$, with $\alpha_i = 0$ otherwise. It measures output correction per unit feature mismatch.

Proposition 3.3 (Local utility of selective supervision)**.** *Suppose the features are differentiable and the output-loss gradient is L_{sm}-Lipschitz on the compared update segments. The additional output-loss reduction $\mathcal{G}(p) = \mathcal{L}_{\text{out}}(\theta_{\text{out}}^+) - \mathcal{L}_{\text{out}}(\theta_p^+)$ satisfies*

$$\mathcal{G}(p) = \delta \sum_i p_i \alpha_i D_i + R_p, \qquad |R_p| \leq \frac{L_{\text{sm}} \delta^2}{2} \|g_p\|_2^2. \tag{4}$$

For two distributions p, q, assume $0 < \alpha_{\min} \leq \alpha_i \leq \alpha_{\max}$ at their positive-mismatch positions. If $\alpha_{\min} S_p > \alpha_{\max} S_q$, then $\mathcal{G}(p) > \mathcal{G}(q)$ for sufficiently small common $\delta > 0$.

The proposition links supervision utility to mismatch and its compatibility with the output objective. When correction efficiencies are positive and comparable, it favors regions whose larger mismatch outweighs the efficiency variation. Combined with the observed concentration in deep layers and early steps, this motivates the selective design in Section 5. For cosine supervision, D_i is the corresponding unit-feature discrepancy (Appendix A.4).

4 Empirical Evidence for the Mismatch Mechanisms

We examine Lemmas 3.1 and 3.2 using output-only DiffusionOPD checkpoint-1000 on SD3.5-M and Z-Image, each paired with PickScore, OCR, and GenEval teachers. We use shared teacher–student inputs for 32 prompts per pair, fixing each prompt's clean latent and noise across $\sigma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

Table 1: Held-out prediction NMSE. Lower is better.

Model	Teacher	t-ctrl.	Scaling
SD3.5-M	PickScore	1.303	**0.748**
	OCR	0.728	**0.547**
	GenEval	5.932	**0.879**
Z-Image	PickScore	1.938	**0.947**
	OCR	3.380	**0.786**
	GenEval	8.243	**0.862**

Evaluating student blocks on their native and complete teacher inputs separates the local discrepancy $b_{\ell,s}$ from the propagated term $p_{\ell,s} = d_{\ell+1,s} - b_{\ell,s}$. We measure image-token energies and their signed interaction (Equation (10)). At the selected deep positions, propagated energy is dominates the discrepancy energy (Figure 3). Upstream state differences thus dominate mismatch at these positions. More results are in Appendix C.

To examine noise dependence, let $\Delta_s = \widehat{y}_S - \widehat{y}_T$ denote the clean-latent estimate difference, with $\star$ marking the reference at $\sigma = 0.9$. We compare the prediction $(\rho_s/\rho_\star)\Delta_\star$ from Lemma 3.2 with the t-control $(t_s/t_\star)\Delta_\star$. Here t_s is the coefficient in $\widehat{y}_M = x_s - t_s \widehat{v}_M$, where $\widehat{v}_M$ is the predicted flow velocity. For each pair, we fit the single parameter τ^2 on 16 prompts and evaluate normalized mean squared prediction error (NMSE) on the remaining 16. Noise-dependent scaling yields lower held-out NMSE in all six pairs (Table 1). This result supports its predictive value beyond the output-conversion factor alone.

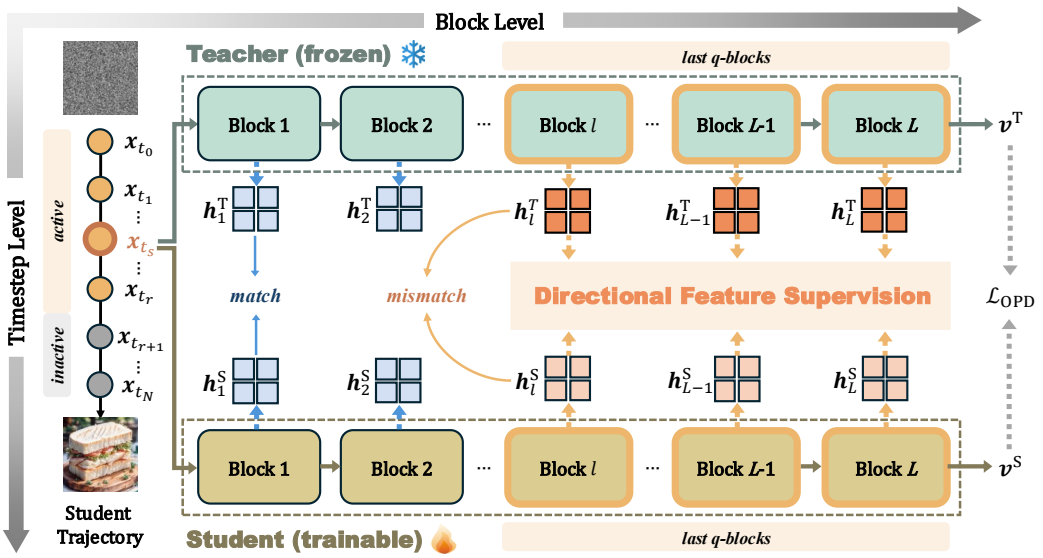


Figure 4: Overview of FocusOPD. Teacher and student are evaluated on shared states from student trajectories. Output supervision follows the base OPD protocol, while directional feature supervision targets deep layers during early denoising. Only the trained student is used at inference.

5 FocusOPD: Structure-Guided Feature Distillation

Motivated by Proposition 3.3, we introduce FocusOPD, a straightforward way to use the persistent depth–temporal mismatch structure in multi-teacher OPD. We concentrate auxiliary supervision on deep layers and early denoising steps, where the observed mismatch is concentrated. The method augments output-level distillation with directional feature alignment in this fixed region (Figure 4).

5.1 Distillation on Student Trajectories

Let S_θ be the student and $\{T_m\}_{m=1}^K$ fixed task-specific teachers sharing its backbone architecture. We sample task m from a distribution π and prompt c from its prompt set $\mathcal{C}_m$. At each round, the student with parameters θ_{roll} generates an N-step trajectory $\tau = (x_{t_0}, \dots, x_{t_N})$, where $t_0 > \cdots > t_N$. Optimization uses detached states $\bar{x}_{t_s} = \text{sg}[x_{t_s}]$ for $s = 0, \dots, N-1$, with sg denoting stop-gradient. Teacher T_m and student S_θ receive the same $(\bar{x}_{t_s}, t_s, c)$. We retain the base output objective $\mathcal{L}_{\text{OPD}}$, including its targets, weights, and query set $\mathcal{I}_{\text{out}} \subseteq \{0, \dots, N-1\}$.

5.2 Selecting the Supervision Region

We select a contiguous deep-layer window and an early denoising window. For a backbone with L blocks, let $q \in \{1, \dots, L\}$ be the number of selected blocks. Their output positions are

$$\mathcal{D}_q = \{L - q + 1, \dots, L\}. \tag{5}$$

For an early-window fraction $r \in [1/N, 1]$, define

$$m_r = \lfloor Nr \rfloor, \qquad \mathcal{I}_{\text{feat}}(r) = \{0, \dots, m_r - 1\}. \tag{6}$$

The resulting region is $\mathcal{D}_q \times \mathcal{I}_{\text{feat}}(r)$, with early steps measured from the high-noise start of each trajectory. Its use as a fixed region is motivated by the observed persistence and Proposition 3.3. Feature queries are specified independently of $\mathcal{I}_{\text{out}}$, so output supervision continues according to the base protocol beyond the feature window.

5.3 Directional Feature Supervision

For $M \in \{S_\theta, T_m\}$, let $h^M_{\ell,s,p} \in \mathbb{R}^{d_h}$ be image token p after ℓ blocks at $(\bar{x}_{t_s}, t_s, c)$, where d_h is the feature dimension and P is the number of image tokens. We extract features from

the text-conditional branch; output predictions use the base guidance configuration. With $\widehat{h} = h/\max(\|h\|_2, \epsilon)$ and norm floor $\epsilon > 0$, the feature loss at step s is

$$\mathcal{L}_{\text{feat}}^{m,s} = \frac{1}{|\mathcal{D}_q|P} \sum_{\mathcal{L}\in\mathcal{D}_q} \sum_{p=1}^{P} \left(1 - \left\langle \widehat{h}_{\ell,s,p}^{S_\theta}, \text{sg}\left[\widehat{h}_{\ell,s,p}^{T_m}\right] \right\rangle\right). \tag{7}$$

This objective aligns feature directions, averaging over selected blocks and image tokens. Its normalized-feature formulation allows Proposition 3.3 to apply directly to cosine-loss gradients (ref Appendix A.4 for details).

The joint training objective is

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{OPD}}(\theta) + \lambda\, \mathbb{E}_{m,c,\tau}\left[\frac{1}{N} \sum_{s\in\mathcal{I}_{\text{feat}}(r)} \mathcal{L}_{\text{feat}}^{m,s}\right], \tag{8}$$

where $\tau \sim S_{\theta_{\text{roll}}}(\cdot \mid c)$ and $\lambda \geq 0$ controls feature supervision. The auxiliary loss is normalized by the full trajectory length N. We keep (q, r, λ) fixed during training and shared across task teachers; concrete settings and supervision-budget comparisons are reported in Section 6.

Training replays states in $\mathcal{I}_{\text{out}} \cup \mathcal{I}_{\text{feat}}(r)$, applying each loss at its designated queries and sharing teacher forward passes where queries overlap. Only student parameters are updated, with fresh student trajectories collected in subsequent rounds. Inference uses the trained student without additional components.

6 Experiments

6.1 Experimental Setup

Models and Evaluation. Following the multi-teacher setting of DiffusionOPD (Li et al., 2026), we evaluate two diffusion backbones, SD3.5-M (Esser et al., 2024) and Z-Image (Z-Image Team et al., 2025), at a resolution of 512×512. For each backbone, the student and three task-specific teachers are initialized from the same base model, with teachers specializing in compositional alignment, text rendering, and human preference optimization. For SD3.5-M, we use the teachers trained in DiffusionOPD, while for Z-Image we train three teachers using the same task-specific RL protocols. All students generate ten-step on-policy trajectories with a deterministic flow solver and only LoRA (Hu et al., 2022) parameters are optimized. All methods are trained for 1000 epochs.

We evaluate compositional alignment on GenEval (Ghosh et al., 2023) and report its six category scores and overall score. Text rendering is evaluated on a held-out OCR split using OCR accuracy. Human preference and visual quality are evaluated on DrawBench (Saharia et al., 2022) with PickScore (Kirstain et al., 2023), HPSv2.1 (Wu et al., 2023), LAION-Aesthetic (Schuhmann, 2022), and ImageReward (Xu et al., 2023). We use the same evaluation prompts, resolution, sampler, guidance settings, random seeds, and evaluator versions for all methods within each backbone.

Baselines. We compare the base model and single-task teachers with multi-teacher OPD methods. DiffusionOPD (Li et al., 2026) and DanceOPD (Zhou et al., 2026) represent output-level distillation, while STEP-OPD (Wei et al., 2026) additionally supervises internal representation dynamics. Our method builds on each OPD objective while preserving its output target and teacher-routing strategy.

6.2 Main Results

Table 2 compares each baseline with and without our feature supervision: DiffusionOPD, DanceOPD, and STEP-OPD are each followed by a corresponding "+ FocusOPD" row. This paired presentation tests whether the proposed feature objective is useful across different OPD output objectives and query strategies. The SD3.5-M and Z-Image experiments use the same teacher roles and evaluation protocol, with depth and denoising windows calibrated per backbone.

Main-result analysis. We compare output-only, uniform KD, and our selective method on a fixed replay bank, pairing depth/denoising mismatch profiles with generation scores. This analysis tests whether better alignment in the supervised regions accompanies better generation quality,

Table 2: Multi-teacher distillation on GenEval, OCR, and DrawBench. The table reports all GenEval categories, OCR accuracy, and four independent DrawBench quality metrics. All metrics are higher-is-better; Scores are reported at checkpoint 1000. Gray rows highlight our method applied to each baseline.

Method	GenEval							OCR	DrawBench			
	Sing.	Two	Cnt.	Col.	Pos.	C-Att.	Ovr.	Acc.	Pick	Aes.	HPS	ImgR
SD3.5-M												
Single-task teacher references												
Base model	0.963	0.889	0.613	0.809	0.250	0.570	0.682	0.675	22.411	5.388	0.281	0.901
GenEval teacher	1.000	1.000	0.950	0.936	0.960	0.900	0.958	0.507	22.061	5.202	0.248	0.611
OCR teacher	0.988	0.838	0.663	0.787	0.310	0.480	0.678	0.961	22.351	5.273	0.273	0.938
Preference teacher	1.000	0.970	0.625	0.894	0.350	0.640	0.746	0.729	24.049	6.237	0.346	1.520
Multi-objective OPD methods												
DanceOPD	0.988	**1.000**	0.900	**0.926**	0.960	0.790	0.927	0.952	23.573	5.872	0.317	1.367
+ **FocusOPD (ours)**	**1.000**	0.990	0.900	0.904	0.960	0.830	0.931	0.945	23.736	5.986	0.326	1.441
STEP-OPD	1.000	1.000	0.975	0.926	0.980	0.900	0.963	0.961	**24.048**	6.255	**0.345**	1.518
+ **FocusOPD (ours)**	**1.000**	**1.000**	0.975	0.915	0.980	**0.930**	0.967	**0.961**	24.025	**6.260**	0.345	**1.520**
DiffusionOPD	**1.000**	**1.000**	0.925	0.915	**0.990**	0.890	0.953	0.958	23.797	6.052	0.333	1.471
+ **FocusOPD (ours)**	**1.000**	**1.000**	**0.988**	**0.926**	**0.990**	0.900	**0.967**	0.960	23.944	6.199	0.340	1.517
Z-Image												
Single-task teacher references												
Base model	0.988	0.889	0.625	0.777	0.380	0.480	0.690	0.779	22.177	5.233	0.270	0.780
GenEval teacher	1.000	0.990	0.913	0.936	0.970	0.880	0.948	0.835	22.287	5.113	0.262	0.893
OCR teacher	0.975	0.838	0.600	0.798	0.360	0.410	0.664	0.876	22.226	5.220	0.272	0.797
Preference teacher	1.000	0.899	0.600	0.819	0.320	0.620	0.710	0.867	23.485	5.942	0.332	1.361
Multi-objective OPD methods												
DanceOPD	0.975	0.828	0.588	0.883	0.430	0.650	0.726	0.873	22.894	5.531	0.300	1.100
+ **FocusOPD (ours)**	0.975	0.960	0.763	0.894	**0.990**	0.800	0.897	0.875	23.203	5.691	0.315	1.247
STEP-OPD	0.988	**0.990**	0.913	**0.968**	0.970	0.850	0.946	0.880	23.547	5.954	**0.334**	1.360
+ **FocusOPD (ours)**	0.988	**0.990**	0.913	0.979	0.970	0.850	0.948	**0.881**	**23.571**	**5.958**	0.334	**1.386**
DiffusionOPD	**1.000**	**1.000**	**0.913**	0.883	0.970	0.810	0.929	0.884	23.467	5.867	0.328	1.342
+ **FocusOPD (ours)**	**1.000**	**1.000**	**0.938**	0.957	0.970	**0.860**	**0.954**	0.881	23.558	5.918	0.333	1.378

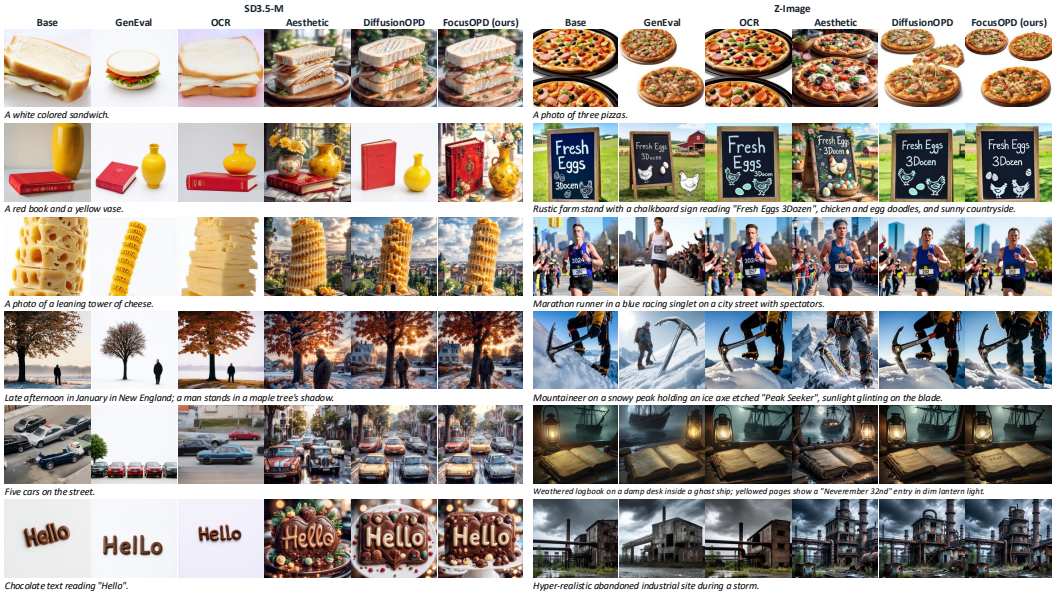


Figure 5: Qualitative comparison of generation results. Our method produces images with better compositional alignment, text rendering quality, and visual aesthetics compared to baseline methods.

and whether the method with the lowest global feature error necessarily has the best task scores. Results confirm that selective supervision in the diagnosed regions (deep layers, early denoising) improves generation more effectively than uniform supervision across all positions.

Cross-baseline and cross-backbone consistency. Our method consistently improves different OPD baselines, as shown in Table 2. DiffusionOPD, DanceOPD, and STEP-OPD each use different output objectives and query strategies, yet all benefit from the same Last-12 Front-40% feature

supervision on both SD3.5-M and Z-Image backbones. This confirms that the proposed feature objective is complementary to various output-distillation protocols.

The same table also shows that the diagnostic-guided selection strategy generalizes from SD3.5-M to Z-Image. While both backbones use the same Last-q principle (deep-layer, early-phase supervision), the specific q and r values are calibrated independently for each architecture through backbone-specific diagnostics and ablations.

6.3 Ablation Studies

All ablation studies use SD3.5-M with DiffusionOPD as the base. We average feature loss over tokens, layers, and active steps. Cosine distance and coefficient λ are fixed unless varied explicitly. Results are reported on the development prompt set with fixed generation seeds.

Table 3: Ablation of feature supervision components on SD3.5-M with DiffusionOPD. Training time per update is measured over steps 11–1000; teacher feature cache is the CPU-stored tensor payload per rank. All feature variants use cosine distance with $\lambda = 0.3$; all methods train the full 24-block transformer.

Configuration	Depth	Time	GenEval	OCR	Pick	HPS	Aes	ImgR	Time/up (s)	Cache (MiB)
Output KD only	–	–	0.951	**0.961**	**23.654**	6.146	0.339	1.502	16.54	0
w/o depth selection	All-24	Front 40%	**0.965**	0.959	23.565	6.208	0.341	**1.505**	19.88	864
w/o temporal selection	Last-12	Full	0.951	0.960	23.556	6.183	0.341	1.500	19.88	2160
All blocks, full time	All-24	Full	0.953	0.958	23.574	6.198	**0.342**	1.501	21.77	2160
FocusOPD	**Last-12**	**Front 40%**	0.960	0.959	23.545	**6.214**	0.340	1.497	**18.99**	**432**

Feature-selection ablation. Table 3 examines whether depth and temporal selection are necessary for effective feature supervision. The output-only baseline establishes that multi-teacher OPD succeeds without auxiliary features (GenEval 0.951, HPS 6.146). Adding feature supervision uniformly across all blocks and all timesteps (All-24/Full) brings modest gains in some metrics (GenEval 0.953, Aes 0.342) but requires the longest training time (21.77s per update) and largest teacher cache (2160 MiB per rank). Comparing the two selection dimensions shows that early temporal coverage is important: supervising the full trajectory gives no GenEval improvement over output-only, whereas Front-40% improves generation quality. For depth, All-24/Front-40% reaches the highest GenEval (0.965) but costs more memory and time than our Last-12/Front-40% setting, which achieves GenEval 0.960 and HPS 6.214 with a smaller cache and faster updates. The detailed depth and temporal trends are analyzed below.

Table 4: Depth window ablation (fixed Front-40% temporal coverage). Block indices are zero-based.

Depth window	GenEval	OCR	Pick	HPS	Aes	ImgR
Output KD only	0.951	0.961	**23.654**	6.146	0.339	1.502
First-6 (0–5)	0.948	0.958	23.652	6.207	**0.342**	**1.513**
Mid-6 (8–13)	0.958	0.959	23.616	6.144	0.340	1.507
Last-6 (18–23)	0.962	**0.962**	23.522	6.199	0.340	1.488
First-12 (0–11)	0.950	0.957	23.629	6.149	0.341	1.506
Mid-12 (6–17)	0.953	0.957	23.607	6.184	0.341	1.506
Last-12 (12–23)	0.960	0.959	23.545	6.214	0.340	1.497
Last-18 (6–23)	0.962	0.959	23.562	6.200	0.341	1.501
All blocks (0–23)	**0.965**	0.959	23.565	6.208	0.341	1.505
First-6+Last-6	0.962	0.959	23.554	**6.225**	0.341	1.498

Table 5: Temporal window ablation (fixed Last-12 depth coverage). Step indices are zero-based.

Temporal window	GenEval	OCR	Pick	HPS	Aes	ImgR
Output KD only	0.951	0.961	**23.654**	6.146	0.339	**1.502**
Front 20% (0–1)	0.959	0.959	23.543	6.174	0.339	1.481
Front 30% (0–2)	0.957	0.960	23.543	6.200	0.340	1.490
Front 40% (0–3)	0.960	0.959	23.545	**6.214**	0.340	1.497
Front 50% (0–4)	0.960	0.960	23.549	**6.214**	**0.341**	1.501
Front 60% (0–5)	0.956	0.960	23.554	6.200	**0.341**	1.500
Full trajectory (0–9)	0.951	0.960	23.556	6.183	**0.341**	1.500
Middle 40% (3–6)	0.953	**0.963**	23.590	6.174	0.339	1.481
Last 40% (6–9)	0.954	0.960	23.579	6.092	0.335	1.484
Front 20%+Last 20%	**0.963**	0.961	23.564	6.198	**0.341**	1.495

Depth selection. Table 4 examines which layers benefit most from feature supervision, holding temporal coverage fixed at Front-40%. Shallow supervision (First-6, blocks 0–5) yields minimal improvement over output-only (GenEval 0.948 vs. 0.951), suggesting that early-layer features are already well-aligned through output distillation. Middle layers (Mid-6, Mid-12) show moderate gains but remain below deep-layer configurations.

Deep-layer supervision consistently outperforms shallow supervision: Last-6 (blocks 18–23) reaches GenEval 0.962 and OCR 0.962, the highest OCR among all depth windows. Expanding to Last-12 (blocks 12–23) achieves the best HPS (6.214) while maintaining strong compositional performance (GenEval 0.960). Further expansion to Last-18 or All-24 increases GenEval slightly

(0.962, 0.965) but at diminishing returns: All-24 requires double the feature computation of Last-12 with only +0.5% GenEval gain. Non-contiguous selection (First-6+Last-6) does not outperform Last-12, confirming that a contiguous deep window suffices. These results support the hypothesis that mismatch concentrated in deeper blocks benefits most from explicit alignment, as these representations are more task-specific and further from the shared backbone prior.

Temporal selection. Table 5 varies the denoising-step coverage with Last-12 depth fixed. Supervising the full trajectory (steps 0–9) provides no gain over output-only (GenEval 0.951 for both), despite supervising 2.5× more layer-step pairs than Front-40%. Middle 40% (steps 3–6) and Last 40% (steps 6–9) similarly underperform, with Last-40% yielding the lowest HPS (6.092) and Aesthetics (0.335) among all temporal configurations.

In contrast, early-phase supervision drives consistent improvements. Front-20% (steps 0–1) already lifts GenEval from 0.951 to 0.959. Extending to Front-40% (steps 0–3) reaches GenEval 0.960 and HPS 6.214, the best preference score across all temporal windows. Front-50% matches this GenEval and HPS, with marginal Aesthetics gain (+0.001), but extending further to Front-60% or Full trajectory degrades performance. The sharp drop from Front-50% (GenEval 0.960) to Full (0.951) indicates that late denoising steps add no value to feature supervision and may even dilute the training signal from high-noise steps. This occurs because noise levels are low and model priors converge at late steps. This empirically confirms the prediction that feature mismatch matters most when noise modulates sensitivity to prior differences, precisely the regime our theory identifies.
Feature-loss formulation. Table 6 compares cosine distance with teacher-energy-normalized MSE under the selected Last-12 and Front-40% windows. Cosine distance gives the strongest GenEval and HPS scores, supporting its use as the default feature loss.

Table 6: Feature loss formulations on SD3.5-M with Last-12 and front-40% coverage. Both feature variants use $\lambda = 0.3$. Normalized MSE divides each sample/layer squared error by teacher mean squared energy, then averages over selected layers; the denominator floor is 10^{-8}. This compares formulations at a fixed coefficient, not matched gradient magnitudes.

Feature loss type	GenEval	OCR	PickScore	HPSv2.1	Aes	ImgRwd
Output KD only (no feature)	0.951	<u>0.961</u>	**23.654**	6.146	<u>0.339</u>	**1.502**
Teacher-energy-normalized MSE	<u>0.955</u>	**0.963**	23.509	<u>6.210</u>	**0.340**	1.491
Cosine distance	**0.960**	0.959	<u>23.545</u>	**6.214**	**0.340**	<u>1.497</u>

7 Related Work

We provide a detailed discussion of related work in Appendix B. On-policy distillation (OPD) trains students on their own trajectories, reducing the distribution mismatch between training and inference (Agarwal et al., 2024). Recent studies (Li et al., 2026; Fang et al., 2026; Zhou et al., 2026; Jiang et al., 2026) extend OPD to multi-teacher capability integration in diffusion models, but primarily rely on output-level supervision. Feature distillation Romero et al. (2015); Lee et al. (2024); Yu et al. (2025); Wei et al. (2026) adds intermediate representation supervision, while selective knowledge distillation (Liang et al., 2023; Liu et al., 2024) adapts the supervised content or location according to teacher relevance. However, how to select feature supervision across network depth and denoising stages in multi-teacher diffusion OPD remains underexplored. Building on output-level OPD, our method supervises conditional image-token representations in selected deep blocks during early denoising to study their depth- and stage-dependent benefits.

8 Conclusion

In this work, we investigate where and when intermediate supervision is most effective for multi-teacher on-policy distillation in diffusion models. We identify a persistent depth–temporal structure in teacher–student mismatch and provide theoretical and empirical evidence linking it to selective supervision. Building on these findings, we introduce FocusOPD, which complements output-level distillation with directional feature alignment in deep layers during early denoising. Experiments across two diffusion backbones and multiple OPD baselines show consistent gains, demonstrating the effectiveness of structure-guided feature supervision for multi-teacher capability integration.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation, polish and clarify the writing, and support the retrieval and discovery of related work. We also used them to support environment construction, testing, and monitoring in our experiments. Specifically, we used large language models (LLMs), including GPT-6 and Claude Opus 5, to improve the clarity and readability of the manuscript. We have reviewed all AI-assisted work and take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

Patrick Esser et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

Zhen Fang, Wenxuan Huang, Yu Zeng, Yiming Zhao, Shuang Chen, Kaituo Feng, Yunlong Lin, Lin Chen, Zehui Chen, Shaosheng Cao, and Feng Zhao. Flow-OPD: On-policy distillation for flow matching models. *arXiv preprint arXiv:2605.08063*, 2026.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Dengyang Jiang, Xin Jin, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Ruoyi Du, Xiangpeng Yang, Qilong Wu, Zhen Li, Peng Gao, Harry Yang, and Steven Hoi. D-OPSD: On-policy self-distillation for continuously tuning step-distilled diffusion models. *arXiv preprint arXiv:2605.05204*, 2026.

Yuval Kirstain et al. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 2023.

Youngwan Lee, Kwanyong Park, Yoorhim Cho, Yong-Ju Lee, and Sung Ju Hwang. KOALA: Empirical lessons toward memory-efficient and fast diffusion models for text-to-image synthesis. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5c4e0e38691e2aa08bba4cefc4c6e852-Abstract-Conference.html.

Quanhao Li et al. DiffusionOPD: A unified perspective of on-policy distillation in diffusion models. *arXiv preprint arXiv:2605.15055*, 2026.

Chen Liang, Simiao Zuo, Qingru Zhang, Pengcheng He, Weizhu Chen, and Tuo Zhao. Less is more: Task-aware layer-wise distillation for language model compression. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20852–20867. PMLR, 2023. URL https://proceedings.mlr.press/v202/liang23j.html.

Kang Liu, Yingyi Zhang, Jingyun Zhang, Jinmin Li, Jun Wang, Shaoming Wang, Chun Yuan, and Rizen Guo. DFD: Distilling the feature disparity differently for detectors. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32421–32430. PMLR, 2024. URL https://proceedings.mlr.press/v235/liu24cd.html.

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

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

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6550.

Chitwan Saharia et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 2022.

Christoph Schuhmann. LAION-Aesthetics. https://laion.ai/blog/laion-aesthetics/, 2022.

Qingyan Wei, Guangzhao Li, Xiaobing Tu, Yinggui Wang, Xiantao Zhang, Jinkui Ren, Xiaohong Liu, and Linfeng Zhang. STEP-OPD: Rethinking output targets and internal dynamics in on-policy distillation for diffusion models. *arXiv preprint arXiv:2608.04887*, 2026.

Xiaoshi Wu et al. Human preference score v2. https://github.com/tgxs002/HPSv2, 2023. Evaluation checkpoint: v2.1.

Jiazheng Xu et al. ImageReward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 2023.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d9e42b4d7163931f3689d6d6fbaa11d0-Abstract-Conference.html.

Z-Image Team et al. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Wei Zhou, Xiongwei Zhu, Zelin Xu, Bo Dong, Lixue Gong, Yongyuan Liang, Meng Chu, Leigang Qu, Lingdong Kong, Wei Liu, and Tat-Seng Chua. DanceOPD: On-policy generative field distillation. *arXiv preprint arXiv:2606.27377*, 2026.

A Proofs of the Mismatch and Supervision Results

A.1 Depth-wise Discrepancy Transport

Proof of Lemma 3.1. Fix s and suppress its index. Adding and subtracting $F_\ell^S(z_\ell^T)$ and applying the fundamental theorem of calculus gives

$$d_{\ell+1} = b_\ell + A_\ell d_\ell, \qquad A_\ell = \int_0^1 J F_\ell^S(z_\ell^T + u d_\ell)\, du, \tag{9}$$

where J denotes the input Jacobian. Recursively substituting this identity gives Equation (1), including the initial embedded-state difference d_0. □

For the diagnostic in Section 4, let $p_\ell = A_\ell d_\ell = F_\ell^S(z_\ell^S) - F_\ell^S(z_\ell^T)$ and let Π select image-token coordinates at the block output. Then

$$\|\Pi d_{\ell+1}\|_2^2 = \|\Pi b_\ell\|_2^2 + \|\Pi p_\ell\|_2^2 + 2\langle \Pi b_\ell, \Pi p_\ell \rangle. \tag{10}$$

The cross term permits reinforcement or cancellation, so the transport identity alone does not imply monotonic growth with depth. Shared noisy latents also need not give $d_0 = 0$ when input embeddings differ.

A.2 Noise-dependent Prior Sensitivity

Proof of Lemma 3.2. Gaussian conditioning yields

$$\widehat{y}_M = \frac{a_s \tau^2}{a_s^2 \tau^2 + \sigma_s^2} x_s + \frac{\sigma_s^2}{a_s^2 \tau^2 + \sigma_s^2} \mu_M(c). \tag{11}$$

At the same (x_s, c), subtracting the two estimates cancels the observation term and gives Equation (2). Since $\rho_s = (1 + \tau^2 \mathrm{SNR}_s)^{-1}$ decreases with SNR, weaker observations preserve more of the prior-mean difference. □

The shared covariance is essential to this scalar response. The lemma models clean-latent estimates; it does not impose the same response on hidden features or velocity predictions. In the flow parameterization $x_t = (1-t)y + t\varepsilon$ with $\widehat{v}_M = (x_t - \widehat{y}_M)/t$, for $t > 0$,

$$\widehat{v}_S - \widehat{v}_T = -\frac{t}{\tau^2(1-t)^2 + t^2} \big(\mu_S(c) - \mu_T(c)\big). \tag{12}$$

This coefficient need not grow monotonically with noise. The following update analysis therefore uses the actual feature loss directly.

A.3 Local Utility of Selective Supervision

Proof of Proposition 3.3. All gradients defining the updates are computed at the same θ. If $D_i = 0$, differentiability gives $g_i = J_{e_i}(\theta)^\top e_i(\theta) = 0$, justifying $\alpha_i = 0$. For $u = \theta_{\mathrm{out}}^+$, the Lipschitz-gradient bound gives

$$\begin{aligned} \mathcal{L}_{\mathrm{out}}(u - \delta g_p) &= \mathcal{L}_{\mathrm{out}}(u) - \delta \langle g_{\mathrm{out}}^+, g_p \rangle + r_p, \\ |r_p| &\le \frac{L_{\mathrm{sm}} \delta^2}{2} \|g_p\|_2^2. \end{aligned} \tag{13}$$

Substitute $g_p = \sum_i p_i g_i$ and $\langle g_{\mathrm{out}}^+, g_i \rangle = \alpha_i D_i$ to obtain Equation (4), with $R_p = -r_p$. For two distributions obeying the compatibility bounds, let $\Delta = \alpha_{\min} S_p - \alpha_{\max} S_q > 0$. Then

$$\mathcal{G}(p) - \mathcal{G}(q) \ge \delta \Delta - \frac{L_{\mathrm{sm}} \delta^2}{2} \big(\|g_p\|_2^2 + \|g_q\|_2^2\big). \tag{14}$$

The right-hand side is positive for $0 < \delta < 2\Delta / [L_{\mathrm{sm}}(\|g_p\|_2^2 + \|g_q\|_2^2)]$ when the denominator is positive, and for any positive δ within the smoothness neighborhood when it is zero. Here η is fixed while the common auxiliary weight controls δ. □

Interpretation of the assumptions. Replay states, teacher targets, and the candidate distributions are fixed only for this comparison. The argument uses ordinary Euclidean gradient steps and requires neither a teacher-matching parameter reference nor a quadratic loss or stable training trajectory. The lemmas motivate the mismatch structure; positive and comparable α_i are an additional local compatibility condition. Equal $\sum_i p_i = \sum_i q_i = 1$ fixes coefficient mass, not gradient norm.

The gradient g^+_{out} is evaluated at the output-only comparison point, while every feature gradient remains evaluated at θ; no sequential recomputation is required by the joint update. Replacing g^+_{out} by g would add an error bounded by $L_{\mathrm{sm}}\eta\delta\|g\|_2\|g_p\|_2$, provided the same smoothness bound also holds between θ and θ^+_{out}. This is first order in δ at fixed η. The result supplies a local allocation principle; persistence of the selected region remains an empirical observation.

A.4 Direct Connection to Cosine Supervision

Let $u^M_k = h^M_k/\|h^M_k\|_2$ be unit-normalized image tokens at position i, and let Q be the total number of tokens across the fixed batch. Stacking their differences as $e_i = Q^{-1/2}\,\mathrm{vec}(U^S - U^T)$ gives the exact identity

$$\frac{1}{2}D_i = \frac{1}{Q}\sum_{k=1}^{Q}\big(1 - \langle u^S_k, u^T_k\rangle\big) = \ell_{\cos,i}. \tag{15}$$

Proposition 3.3 thus applies to the actual cosine-loss gradients, without a feature-response linearization. Its mismatch variable is angular discrepancy; a raw-MSE ranking does not by itself establish the same ordering. The identity assumes nonzero tokens with the numerical norm floor inactive. Where the safeguarded loss is differentiable but this identity does not hold, the general gain expansion still applies directly to its loss gradient through $\delta\sum_i p_i\langle g^+_{\mathrm{out}}, \nabla\ell_i\rangle + R_p$.

B Full Related Work.

On-policy distillation for capability integration. On-policy distillation (OPD) (Agarwal et al., 2024) transfers teacher knowledge on student-generated trajectories, addressing the distribution mismatch between training and inference. Recent work (Ma et al., 2026; Li et al., 2026; Fang et al., 2026) extends this paradigm to capability integration from multiple teachers in language and diffusion models. DiffusionOPD Li et al. (2026) derives transition-level supervision for diffusion processes, while Flow-OPD Fang et al. (2026) combines task-specialized teachers with a task-agnostic anchor for multi-task alignment. DanceOPD Zhou et al. (2026) formulates capability transfer as generative field distillation, using sample-wise routing and a single low-noise velocity query. D-OPSD Jiang et al. (2026) applies on-policy self-distillation to continual adaptation of few-step models. Building on output-level OPD, our work investigates intermediate teacher supervision for multi-task capability transfer.

Feature alignment in diffusion models. Feature distillation supplements output matching with intermediate teacher representations (Romero et al., 2015). In diffusion models, KOALA Lee et al. (2024) studies self-attention distillation for compact text-to-image generators, while REPA Yu et al. (2025) aligns denoising features with pretrained visual representations to improve diffusion-transformer training. STEP-OPD (Wei et al., 2026) introduces representation-change alignment alongside output extrapolation, matching the direction and magnitude of inter-block updates. Our method directly aligns conditional image-token representations within selected deep blocks and an early denoising window, with feature supervision scheduled independently of output queries.

Selective and adaptive knowledge distillation. Selective distillation accounts for differences in the relevance and transferability of teacher knowledge. TED Liang et al. (2023) learns task-aware filters for intermediate representations in language-model compression. DFD Liu et al. (2024) assigns different distillation constraints to spatial regions according to teacher–student feature disparity in object detection. SAKD learns a routing policy to adapt distillation locations to individual samples and training iterations. These methods select feature content, spatial constraints, or distillation locations in discriminative tasks. We study which blocks and denoising stages benefit from additional feature supervision in diffusion OPD.

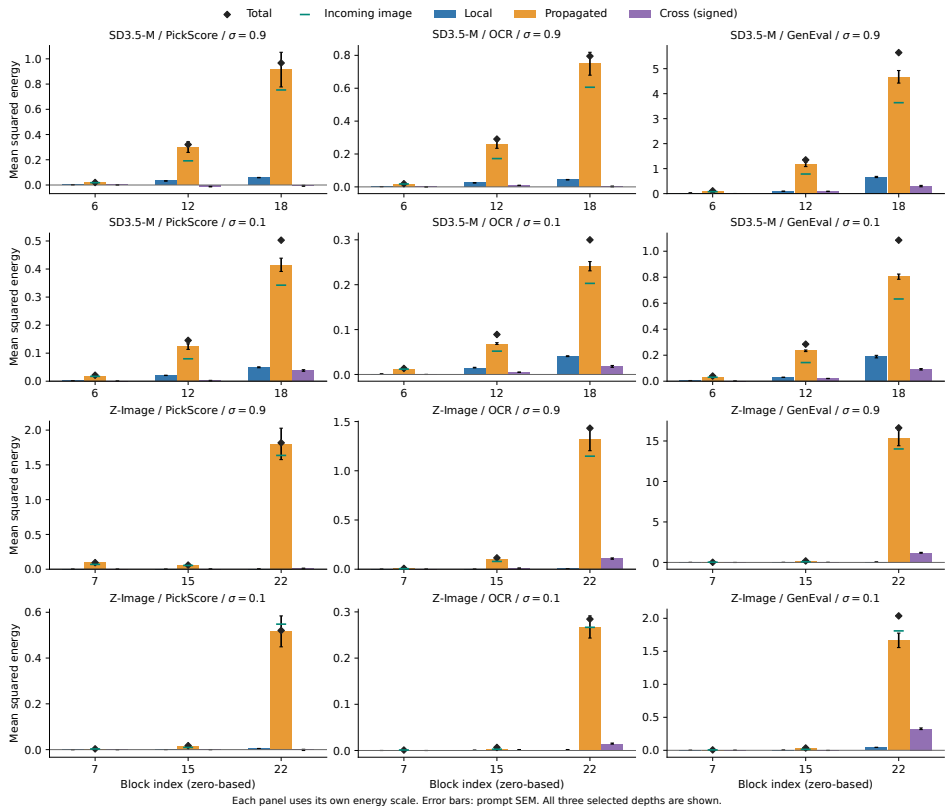


Figure 6: Deep-block decomposition across three teachers for all tested σ. Terms are normalized by total mean discrepancy energy. Block indices are 18 for SD3.5-M and 22 for Z-Image. Error bars: 95% prompt-bootstrap intervals.

C Full results of Empirical Evidence for the Mismatch Mechanisms

Figure 6 presents the full version of Figure 3. As shown, the cumulative effect dominates the teacher–student mismatch across depths, stages, and configurations of different networks.