RETHINKING ON-POLICY DISTILLATION IN DIFFUSION MODELS: PHENOMENOLOGY AND EMPIRICAL INSIGHTS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) consolidates diffusion-model specialists into a shared student by supervising its own denoising trajectories. Yet its mechanisms remain underexplored: What does OPD transfer? What determines transfer at student-visited states? When is distillation worthwhile relative to direct joint optimization? We study three representative OPD methods on two backbones and thirteen routes through budget-controlled comparisons, held-out behavioral evaluation, frozen-trajectory interventions, temporal supervision and teacher-source experiments, and comparisons with direct joint reinforcement learning. First, OPD consolidates specialist gains selectively: DanceOPD leads three of four equal-time settings, whereas DiffusionOPD leads three of four equal-data settings, with winners changing in 17 of 52 matched route comparisons. In the Z-Image two-interface equal-data setting, held-out image-editing and multi-image scores improve by approximately 3.4% and 5.8% over the exact initialization. Second, transfer depends on timing and teacher source: query-matched disagreement weighting improves held-out image editing and multi-image generation by approximately 10.3% and 2.8% over uniform weighting, yet fixed low-noise supervision leads on ten of thirteen routes. Moreover, a reinforcement-learning teacher that scores 6.1% above its supervised-fine-tuning counterpart produces a student that scores 2.9% lower in the design case. Third, under UnifiedReward, direct joint reinforcement learning improves general text-to-image, general image-editing, and held-out image-editing scores by approximately 5.5%, 6.6%, and 93.3% over initialization. Equal-data DiffusionOPD is 2.6% better than direct joint reinforcement learning on general image editing but 16.9% and 24.1% worse on general text-to-image and held-out image editing, respectively. We hope these findings inform future OPD design and use by the community. All resources will be publicly released, including code, checkpoints and datasets.

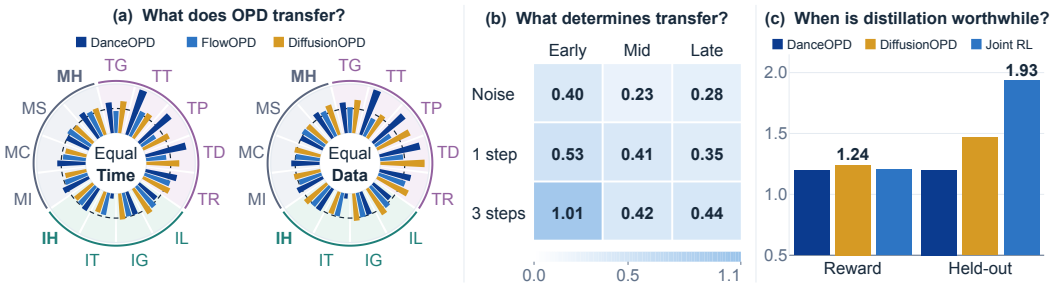

Figure 1. **Overview of three levels of on-policy distillation in diffusion models.** (a) Method rankings in different tasks. (b) Different interventions yield different editing gains. (c) OPD can effectively assimilate training tasks, yet held-out performance remains inferior to that of joint RL.

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.

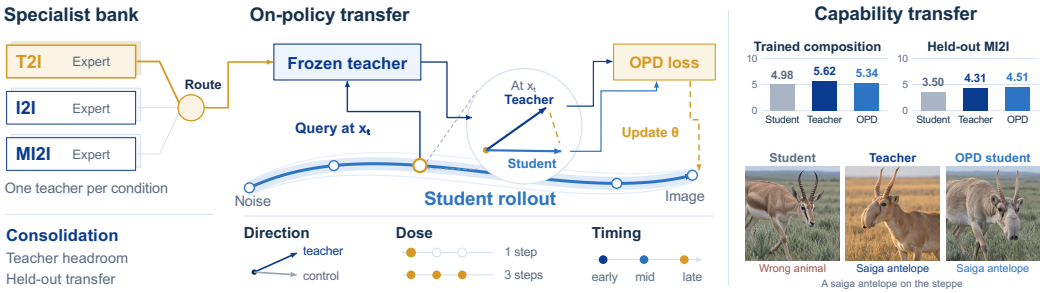

Figure 2. **Overview of on-policy capability consolidation.** Left shows routed supervision on student trajectories; middle shows direction, dose, and timing; right compares student, teacher, and DanceOPD on trained (bottom row) and held-out tasks (top row).

1 Introduction

Diffusion models increasingly support text-to-image generation, image editing, and multi-image composition within a single system (Ho et al., 2020; Rombach et al., 2022; Esser et al., 2024; Cai et al., 2025). Supervised fine-tuning (SFT) and reinforcement learning (RL) can improve individual routes while degrading others, motivating the consolidation of separately acquired specialist capabilities. Thus, on-policy distillation (OPD) (Hinton et al., 2015; Agarwal et al., 2024) is proposed to address this problem by querying specialists along a shared student's own denoising trajectories. (Zhou et al., 2026b;a; Li et al., 2026b; Fang et al., 2026).

However, the mechanism and practical value of OPD remain underexplored. A specialist's own performance does not determine what it can teach a particular student, while teacher resemblance does not necessarily imply better task behavior; supervision timing, teacher source, and training budget further affect the outcome. Prior work has documented teacher–student capacity and compatibility effects in knowledge distillation (Cho & Hariharan, 2019; Mirzadeh et al., 2020; Busbridge et al., 2025; Li et al., 2025), while recent OPD research motivates evaluating transfer relative to the receiving student (Li et al., 2026c). Accordingly, we organize our study around three questions: What does OPD transfer? What determines transfer at student-visited states? When is distillation worthwhile relative to direct joint optimization?

We study three representative OPD methods on SD3.5-M and Z-Image across thirteen text-to-image, image-editing, and multi-image routes. Our study combines budget-controlled comparisons, held-out behavioral evaluation, frozen-trajectory interventions, temporal supervision and teacher-source experiments, and comparisons with direct joint reinforcement learning. Equal-time and equal-data endpoints separate time efficiency from data efficiency, while evaluations relative to the exact initialization distinguish acquired gains from teacher resemblance. Figures 1 and 2 summarize the questions and experimental design.

First, OPD consolidates specialist gains selectively rather than preserving the full task profile automatically. DanceOPD leads three of four equal-time settings, whereas DiffusionOPD leads three of four equal-data settings, with winners changing in 17 of 52 matched route comparisons. In the Z-Image two-interface equal-data setting, held-out image-editing and multi-image scores improve by approximately 3.4% and 5.8% over the exact initialization. Students also exceed their corresponding teachers in 72.2% of 36 eligible held-out cells, showing that teacher-relative gains, initialization-relative gains, and teacher resemblance capture different aspects of transfer.

Second, transfer depends on when supervision is applied and which teacher provides it. Across 480 cases and 4320 frozen-trajectory branches, teacher-directed corrections outperform equal-norm random perturbations in all three task families, but their gains depend on intervention dose and timing. During training, query-matched disagreement weighting improves held-out image editing and multi-image generation by approximately 10.3% and 2.8% over uniform weighting, yet fixed low-noise supervision leads on ten of thirteen routes. Teacher quality is likewise student-relative: in the design case, a reinforcement-learning teacher that scores 6.1% above its supervised-fine-tuning counterpart produces a distilled student that scores 2.9% lower.

Third, the value of distillation depends on the required task profile. Under UnifiedReward, direct joint reinforcement learning improves general text-to-image, general image-editing, and held-out

image-editing scores by approximately 5.5%, 6.6%, and 93.3% over initialization. Equal-data DiffusionOPD is 2.6% better than direct joint reinforcement learning on general image editing but 16.9% and 24.1% worse on general text-to-image and held-out image editing, respectively. Distillation can therefore better retain native rewards, whereas direct joint reinforcement learning preserves broader cross-interface task behavior in the tested settings.

Together, these results characterize OPD as student-relative specialist reuse rather than generic teacher copying. Its value should be assessed by transfer to the receiving student, preservation of the required task profile, and total pipeline cost. We hope these findings provide a principled foundation for understanding when and how on-policy distillation transfers specialist capabilities, guiding the design of more effective strategies for consolidating and optimizing diffusion models.

2 Problem formulation

2.1 Notation

Let $x = (p, I^{\mathrm{src}}, \mathbf{I}^{\mathrm{ref}})$ denote a condition, where p is an instruction, I^{src} is an optional source image, and $\mathbf{I}^{\mathrm{ref}}$ is an ordered, possibly empty tuple of reference images with their roles. Missing image inputs are represented by $\emptyset$. Let $\mathcal{D}$ denote the training-condition distribution and $\mathcal{R}$ the set of specialist training routes. The routing rule $r(x) \in \mathcal{R}$ assigns one frozen teacher $T_{r(x)}$ to the entire trajectory. Each teacher is acquired through supervised fine-tuning or reinforcement learning (Zheng et al., 2026; Black et al., 2024; Prabhudesai et al., 2023; Zhou et al., 2026a).

Let S_θ be the student with trainable parameters θ and fixed initialization S_0. For OPD method m, the current student generates a denoising trajectory

$$\tau = (z_0, \dots, z_K) \sim p_\theta^{(m)}(\cdot \mid x),$$

where $t_0 > \cdots > t_K$ is the noise-time schedule and z_k is the sampler state at t_k, including the current target latent and any solver history required by method m. The distribution $p_\theta^{(m)}$ includes the initial noise and method-specific sampler stochasticity.

At each state z_k, for $k \in \{0, \dots, K-1\}$, define the student and routed-teacher velocity predictions as

$$v_{\theta,k} \coloneqq v_\theta(z_k, t_k, x), \qquad u_k \coloneqq v_{T_{r(x)}}(z_k, t_k, x).$$

Both predictions are evaluated at the same student-visited state and act on the same target latent.

2.2 On-policy distillation

Method m queries the teacher at indices $\mathcal{I}_m(\tau, x) \subseteq \{0, \dots, K-1\}$. Its OPD objective is

$$\mathcal{L}_{\mathrm{OPD}}^{(m)}(\theta) = \mathbb{E}_{\substack{x \sim \mathcal{D} \\ \tau \sim p_\theta^{(m)}(\cdot \mid x)}} \left[\ell_m(\theta; \tau, x, \{(k, u_k) : k \in \mathcal{I}_m(\tau, x)\}) \right],$$

where random query selection is included in the expectation. The trajectory loss ℓ_m retains the method's native supervision, weighting, and regularization. The shared property is that teacher predictions are queried at states visited by the current student (Agarwal et al., 2024; Ross et al., 2011). Rollout differentiation and gradient estimation remain method-specific.

The three methods instantiate this objective differently. DanceOPD matches velocity predictions at sparse student-visited states:

$$\mathcal{L}_{\mathrm{OPD}}^{(\mathrm{Dance})}(\theta) = \mathbb{E}_{\mathrm{Dance}} \left[\sum_{k \in \mathcal{I}_{\mathrm{Dance}}} a_k \, \mathrm{MSE}(v_{\theta,k}, u_k) \right],$$

where $a_k \geq 0$ denotes its native query weighting and normalization (Zhou et al., 2026b). DiffusionOPD matches the means of denoising transitions:

$$\mathcal{L}_{\mathrm{OPD}}^{(\mathrm{Diffusion})}(\theta) = \mathbb{E}_{\mathrm{Diffusion}} \left[\sum_{k \in \mathcal{I}_{\mathrm{Diffusion}}} b_k \left\| \mu_{\theta,k}(z_k, x) - \mu_{T_{r(x)},k}(z_k, x) \right\|_2^2 \right],$$

Change vs. initialization (%) ≤ -20 0 ≥ +20 | 0.88 Best OPD per budget | Equal time | Equal data

Setting	Budget	Method	T2I General	T2I Text	T2I Portrait	T2I Design	T2I Realism	I2I Local	I2I Global	I2I Text	I2I Held-out	MI2I Identity	MI2I Comp.	MI2I Spatial	MI2I Held-out	Wins of 13	Above init. of 13
		Student init.	0.73	0.18	0.57	0.69	0.90	5.16	6.55	0.63	0.60	4.07	4.55	1.23	3.20	-	-
SD3.5-M T2I + I2I	⏱	Dance	**0.81**	**0.32**	**0.64**	0.64	0.96	**5.44**	**6.81**	0.61	0.52	**4.29**	4.45	**1.26**	3.06	**7**	**8**
	⏱	Flow	0.79	0.22	0.57	**0.75**	0.93	5.41	6.65	0.67	**0.68**	4.13	**4.53**	1.16	**3.12**	**4**	**9**
	⏱	Diffusion	0.80	0.15	0.60	0.74	**0.96**	5.25	6.50	**0.68**	0.54	4.08	4.53	1.10	3.08	**2**	**7**
	🗄	Dance	**0.81**	0.32	0.64	0.64	0.96	5.44	**6.81**	0.61	0.52	**4.29**	4.45	**1.26**	3.06	**4**	**8**
	🗄	Flow	0.73	0.24	0.55	0.67	0.95	**5.58**	6.70	**0.68**	**0.58**	4.21	4.40	1.18	3.10	**3**	**7**
	🗄	Diffusion	0.79	**0.36**	**0.70**	**0.77**	**0.98**	5.24	6.68	0.63	0.56	4.16	**4.48**	1.22	**3.21**	**6**	**9**
		Student init.	0.82	0.27	0.65	0.66	1.00	5.19	6.43	0.77	0.78	4.95	4.98	2.51	3.50	-	-
SD3.5-M T2I + I2I + MI2I	⏱	Dance	0.88	0.27	0.64	**0.80**	1.02	5.52	**6.74**	0.89	0.60	**5.22**	**5.34**	4.18	4.50	**4**	**11**
	⏱	Flow	0.82	0.33	0.64	0.76	1.00	5.67	6.61	**1.00**	**0.94**	5.16	5.16	4.12	**4.52**	**3**	**11**
	⏱	Diffusion	**0.88**	**0.37**	**0.70**	0.72	**1.02**	**5.90**	6.54	0.82	0.50	5.15	5.16	**4.50**	4.47	**6**	**12**
	🗄	Dance	**0.88**	0.27	0.64	**0.80**	1.02	5.52	**6.74**	0.89	0.60	5.22	**5.34**	4.18	**4.50**	**5**	**11**
	🗄	Flow	0.83	0.34	0.63	0.73	0.95	5.48	6.36	**0.98**	**0.98**	5.11	5.22	4.30	4.29	**2**	**10**
	🗄	Diffusion	0.87	**0.45**	**0.70**	0.69	**1.02**	**5.69**	6.72	0.79	0.56	**5.26**	5.23	**4.54**	4.42	**6**	**12**
		Student init.	0.96	0.80	0.81	0.86	0.97	6.75	7.84	4.43	4.14	5.31	4.97	2.46	4.46	-	-
Z-Image T2I + I2I	⏱	Dance	1.01	**0.95**	**0.92**	**0.98**	**1.03**	6.65	7.83	3.73	**4.19**	**5.49**	**5.13**	2.47	**4.67**	**8**	**10**
	⏱	Flow	0.94	0.82	0.76	0.83	0.99	6.90	7.82	**4.36**	4.12	5.28	4.91	**2.51**	4.56	**2**	**5**
	⏱	Diffusion	**1.02**	0.83	0.84	0.92	1.01	**7.04**	**7.95**	4.33	4.10	5.19	4.85	2.46	4.59	**3**	**8**
	🗄	Dance	1.01	**0.95**	**0.92**	0.98	1.03	6.65	7.83	3.73	4.19	**5.49**	**5.13**	2.47	4.67	**4**	**10**
	🗄	Flow	0.97	0.77	0.77	0.85	1.00	6.89	7.87	**4.41**	4.18	5.28	4.97	**2.50**	4.58	**2**	**8**
	🗄	Diffusion	**1.03**	0.82	0.89	**1.00**	**1.05**	**7.10**	**8.03**	4.39	**4.28**	5.42	5.05	2.49	**4.72**	**7**	**12**
		Student init.	1.01	1.02	0.84	0.85	0.97	7.03	7.74	3.83	4.43	6.50	5.86	4.63	5.27	-	-
Z-Image T2I + I2I + MI2I	⏱	Dance	1.02	**1.01**	**0.94**	**0.92**	1.04	**7.16**	**8.04**	**4.04**	4.22	6.28	5.52	**4.69**	5.05	**7**	**8**
	⏱	Flow	1.02	0.93	0.81	0.90	1.01	6.99	7.78	3.74	4.37	6.49	**5.81**	4.60	**5.25**	**2**	**4**
	⏱	Diffusion	**1.04**	0.96	0.87	0.84	**1.05**	7.09	7.90	4.00	**4.46**	**6.51**	5.78	4.62	5.15	**4**	**8**
	🗄	Dance	1.02	**1.01**	**0.94**	**0.92**	1.04	**7.16**	**8.04**	4.04	4.22	6.28	5.52	**4.69**	5.05	**6**	**8**
	🗄	Flow	**1.03**	0.94	0.88	0.85	1.04	6.95	7.75	3.63	4.34	**6.49**	**5.77**	4.64	**5.13**	**4**	**6**
	🗄	Diffusion	1.02	0.90	0.92	0.91	**1.11**	7.10	7.99	**4.09**	**4.39**	6.49	5.66	4.52	5.03	**3**	**7**

Table 1. **Comparison of OPD methods across** 13 **tasks on** 2 **backbones.**

where $b_k \geq 0$ is its native weighting, and $\mu_{\theta,k}$ and $\mu_{T_{r(x)},k}$ are the student and teacher means for the next target latent under the method's transition parameterization (Li et al., 2026b). FlowOPD instead uses stochastic student transitions together with its native policy and manifold-regularization objectives (Fang et al., 2026). Thus, the methods share student-state supervision but retain different query densities, rollout distributions, losses, and regularization.

2.3 Evaluation and Time Conventions

A teacher's own-output score measures endpoint behavior, not its usefulness to the receiving student. We therefore evaluate transfer relative to the exact initialization S_0 and separately report behavior on routes outside the training objectives. Teacher-relative gains, initialization-relative gains, and teacher resemblance are treated as distinct quantities.

We also distinguish noise time from denoising progress. Let y_0 be a clean target latent and let ϵ be standard Gaussian noise of the same shape. Under $y_t = (1-t)y_0 + t\epsilon, \quad t \in [0,1]$, larger t denotes a noisier target latent. For a K-step trajectory, we define denoising progress as $\rho_k := \frac{k}{K}$. Sampling proceeds from high to low noise time while ρ_k increases. Early interventions therefore have small ρ_k, whereas low-noise training queries occur at large ρ_k. An inference intervention changes one fixed trajectory; a training query updates the student used to generate subsequent trajectories. Comparison controls and evaluation protocols are given in Appendices B.3 and C.

3 What Does On-Policy Distillation Transfer?

The first question is what does OPD transfer across tasks and interfaces? We answer this question through two experiments. First, Section 3.1 tests whether a shared student retains specialist gains under equal-time and equal-data budgets. Second, Section 3.2 tests whether held-out task behavior follows teacher resemblance.

3.1 Budget-Dependent Consolidation

We audit paired SFT and RL specialists using their optimized route and the cross-route preservation requirement in Appendix B.5, then compare DanceOPD, DiffusionOPD, and FlowOPD on two mix-

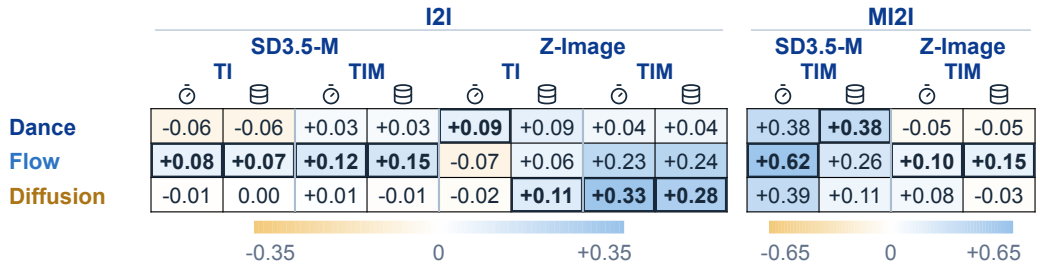

Table 2. **Comparison of held-out gains over the teacher.** Cells report student-minus-teacher changes in mean per-image minimum rubric scores.

tures for both backbones (Zhou et al., 2026b; Li et al., 2026b; Fang et al., 2026). TI combines the general T2I and I2I training routes, whereas TIM additionally includes MI2I composition. Equal-time endpoints use the same elapsed training time; equal-data endpoints use the same condition exposure. A route win is the highest native score among the three methods for a fixed backbone, mixture, budget, and route, and need not exceed the exact initialization S_0. Scores are compared only within their native routes; endpoint and replication details are given in Appendix B.3.

(i) Specialist performance does not establish transfer. Reinforcement learning improves the optimized route in 72.7% of paired specialist comparisons, yet only one endpoint satisfies the predefined cross-route preservation requirement. We therefore evaluate the complete task profile rather than infer transfer from a specialist's own-route score.

(ii) The training budget changes the winning method. DanceOPD wins the most routes in three of four equal-time settings, whereas DiffusionOPD does so in three of four equal-data settings. Among 52 matched backbone, mixture, and route comparisons, 17 change winners between budgets, including eight changes from DanceOPD to DiffusionOPD. In the Z-Image TI setting, DanceOPD's route wins decrease from eight under equal time to four under equal data, while DiffusionOPD's increase from three to seven. Figure 6 reports all winner transitions.

(iii) OPD consolidates specialist gains selectively. DiffusionOPD improves twelve of thirteen routes over S_0 for the SD3.5-M TIM setting under both budgets and for the Z-Image TI setting under equal data. In the latter setting, held-out image editing and multi-image generation improve by approximately 3.4% and 5.8%, respectively. However, FlowOPD improves both held-out suites under both budgets in the SD3.5-M TIM setting despite winning fewer routes overall, and no method consistently improves both held-out suites across budgets in the Z-Image TIM setting. Broad initialization-relative gains and held-out preservation can therefore favor different endpoints.

These comparisons establish a budget-dependent ranking rather than an isolated effect of query density, because the methods also differ in rollout distributions, losses, and regularization. Likewise, TI and TIM use different initializations and training mixtures, so their difference cannot be attributed solely to the added MI2I teacher. The evidence supports selective consolidation, not a universally superior objective or guaranteed held-out preservation.

3.2 Held-out behavior beyond teacher resemblance

Does held-out task behavior follow teacher? To test combined task constraints, we reuse the existing rubric judgments on $n_e = 100$ examples for each held-out route e. For model f, example i, and criterion j, let $q_{eij}(f)$ be the criterion score and let J_e be the number of criteria. We summarize each model by the mean per-image minimum $M_e(f) := \frac{1}{n_e}\sum_{i=1}^{n_e} \min_{1\le j\le J_e} q_{eij}(f)$. This statistic penalizes each image's weakest criterion without introducing a new judge or binary threshold. For cells with a supported teacher interface, we compare $M_e(S_\theta)$ with $M_e(T_r)$. The 36 eligible cells span method, budget, and held-out route; MI2I teacher comparisons are undefined for TI.

(i) Students often exceed their teachers on combined held-out constraints. Students obtain a positive teacher-relative change in 72.2% of the 36 eligible cells, with nine regressions and one tie. The positive-cell rates are 66.7% for DanceOPD, 91.7% for FlowOPD, and 58.3% for DiffusionOPD.

(ii) The held-out metric changes the ranking of DanceOPD and DiffusionOPD. At the individual-criterion level, the rates of positive student–teacher changes are 58.8% for FlowOPD, 44.3% for DiffusionOPD, and 38.5% for DanceOPD. When we instead take each image's lowest criterion score before averaging within a held-out cell, the proportions of cells with positive teacher-

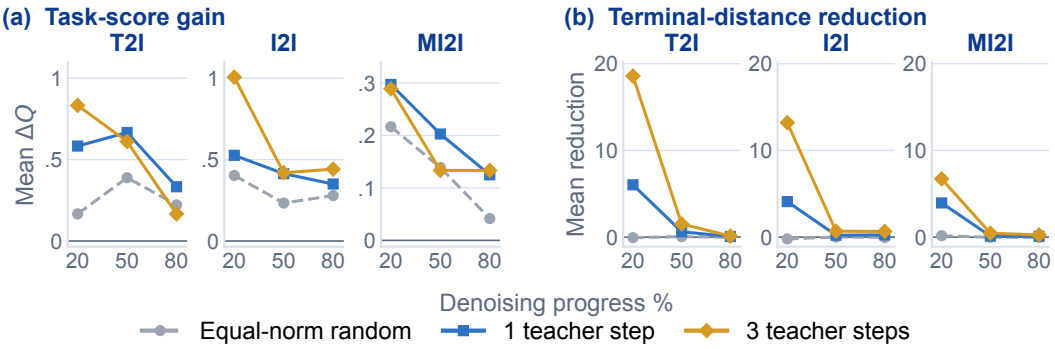

Figure 3. **Effects of trajectory interventions.** (a) Mean native-rubric task gains from random, one-step, and three-step interventions across cases. (b) Mean terminal-distance reductions toward the full teacher under the same interventions.

relative changes are 91.7%, 66.7%, and 58.3% for FlowOPD, DanceOPD, and DiffusionOPD, respectively. FlowOPD remains first, but DanceOPD moves ahead of DiffusionOPD. These rates describe different evaluation units and should not be compared directly; neither measure alone establishes improvement over S_0 or success on every example.

We assess teacher resemblance with linear centered kernel alignment on matched conditions, sampler states, and network depths, reporting $\Delta\text{CKA} := \text{CKA}(S_\theta, T_r) - \text{CKA}(S_0, T_r)$ (Kornblith et al., 2019). DINOv2 patch correspondence separately measures output-to-reference and student-to-teacher visual-feature similarity under the fixed protocol in Appendix F.2 (Oquab et al., 2023).

(iii) Teacher alignment is localized and reference-dependent. For TI, S_0 is itself the image-editing teacher, so the baseline teacher CKA is maximal and ΔCKA is necessarily non-positive. This setting therefore measures how much the student preserves the initial teacher representation rather than whether it converges toward a distinct teacher. For TIM, 30.6% of the changes are positive and concentrate at particular sampler states and network depths (Figure 8a). Representation alignment is therefore a localized diagnostic whose interpretation depends on the reference model and measurement location; direct task judgments remain necessary for evaluating held-out transfer.

(iv) Patch correspondence does not reproduce the held-out task ordering. DINOv2 correspondence shows only small between-method differences and yields a different ordering from the held-out criterion gains. In matched examples, similar student-to-teacher correspondence can accompany subject loss, while stronger source-to-output correspondence can accompany an incomplete replacement in Figure 9. Appendix F.2 reports the corresponding details.

4 What determines transfer?

The second question is what determines transfer at student-visited states? Section 4.1 tests correction direction, dose, and timing on frozen student trajectories. Sections 4.2 and 4.3 then test temporal supervision and teacher-source transferability during training (Salimans & Ho, 2022; Song et al., 2023).

4.1 Correction and continuation on frozen trajectories

For a fixed condition x, sampler, initial noise, checkpoint, step count, classifier-free guidance, and sampler randomness, let $\Phi_k(z_k, v)$ denote one complete state update using prediction v, including any required solver history. The unmodified student takes $z_{k+1} = \Phi_k(z_k, v_{\theta,k})$, while replacing its prediction with the routed teacher injects the correction $\delta_k := \Phi_k(z_k, u_k) - \Phi_k(z_k, v_{\theta,k})$. All later steps use the original student, so the experiment measures how a teacher-directed correction is propagated by student continuation.

For illustration, we examine one teacher step at approximately 20%, 50%, and 80% denoising progress on three preselected large-gap failures. Our aggregate diagnostic uses 480 cases (180 text-to-image, 180 image-editing, and 120 multi-image cases), without conditioning case inclusion on the teacher–student score gap or student failure status. At each of the three positions, we compare an equal-norm random-direction control, one teacher step, and a centered three-step teacher window, producing 4320 scored branches. For each case, ΔQ is the branch's native-rubric task score mi-

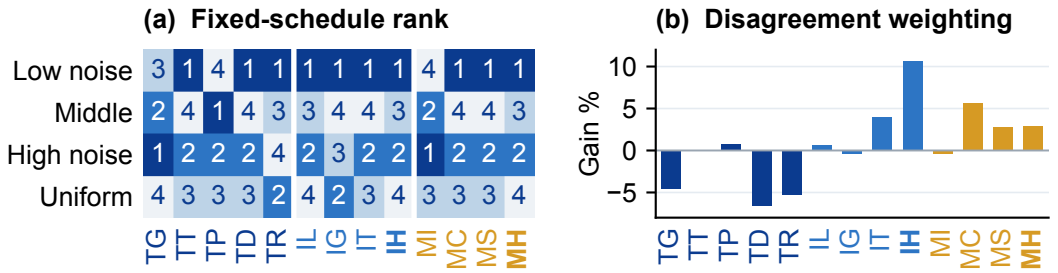

Figure 4. **Effects of temporal supervision.** (a) Low-noise supervision ranks first on ten routes among fixed schedules. (b) Disagreement weighting improves eight routes over query-matched uniform weighting, including both held-out suites.

nus the unmodified student's score. We separately report the reduction in Euclidean terminal latent distance to the full teacher, where a positive value means movement toward the teacher.

(i) Correction direction matters. Teacher-directed single-step corrections produce larger mean task gains than equal-norm random controls: 0.528 versus 0.259 for T2I, 0.430 versus 0.305 for I2I, and 0.208 versus 0.132 for MI2I, all relative to the unmodified student. They also reduce the final latent distance to the full teacher by 2.246, 1.530, and 1.394 in the three task families, respectively, whereas the random controls yield approximately zero distance reduction.

(ii) The useful dose and timing depend on the task family. Across all the cases, early three-step windows give the largest mean gains for text-to-image generation and image editing, whereas an early single step is slightly better for multi-image generation. The three-step-minus-one-step differences are 0.009, 0.193, and -0.023 for T2I, I2I, and MI2I, respectively. Early three-step windows also produce the largest terminal-distance reduction, so trajectory proximity and task gain do not share the same ordering. Figures 10 and 11 report the outcome distributions and six matched editing cases.

(iii) Maximum disagreement is not a better timing rule than fixed early intervention. At a tested state, prediction disagreement is $d_k := \|v_{\theta,k} - u_k\|_2$. Selecting the largest d_k among the three positions identifies the best tested one-step timing in 68.8% of cases and selects an early position in 91.7%. However, it does not improve any task-family mean over fixed early intervention: the T2I and I2I means are tied, while the MI2I mean decreases from 0.297 to 0.281 (Figure 8b). We therefore test disagreement as a training weight rather than as an inference-time timing rule.

4.2 Temporal Supervision

The temporal-supervision experiment uses the same SD3.5-M I2I-GLOBAL RL initialization, frozen MI2I-COMPOSITION RL teacher, optimizer, $3,000$ General-I2I conditions, and 94 updates. Four fixed schedules query low-, middle-, high-noise, or uniformly sampled states. A query-matched pair instead uses the same index set $\mathcal{I}_{\mathrm{DW}}$ at 20%, 50%, and 80% denoising progress, differing only in the weights applied to the three queries. The disagreement-weighted loss is $\ell_{\mathrm{DW}}(\theta;\tau,x) = \sum_{k\in\mathcal{I}_{\mathrm{DW}}} w_k \,\mathrm{MSE}(v_{\theta,k}, u_k)$, with $w_k = \mathrm{stopgrad}\big(d_k / \sum_{h\in\mathcal{I}_{\mathrm{DW}}} d_h\big)$. If the denominator is at most 10^{-12}, all weights are $1/3$; the uniform control always uses $w_k = 1/3$.

(i) Fixed low-noise supervision gives the strongest overall task profile. It ranks first on ten of thirteen routes, including both held-out suites. Early inference interventions and low-noise training need not prefer the same time because the former changes one fixed trajectory, whereas the latter changes the student used for future trajectories.

(ii) Disagreement weighting improves the query-matched uniform baseline but not fixed low-noise supervision. It improves eight of thirteen routes over uniform weighting, including relative gains of approximately 10.3% on held-out image editing and 2.8% on held-out multi-image generation. It also causes regressions and provides no overall advantage over the fixed low-noise schedule.

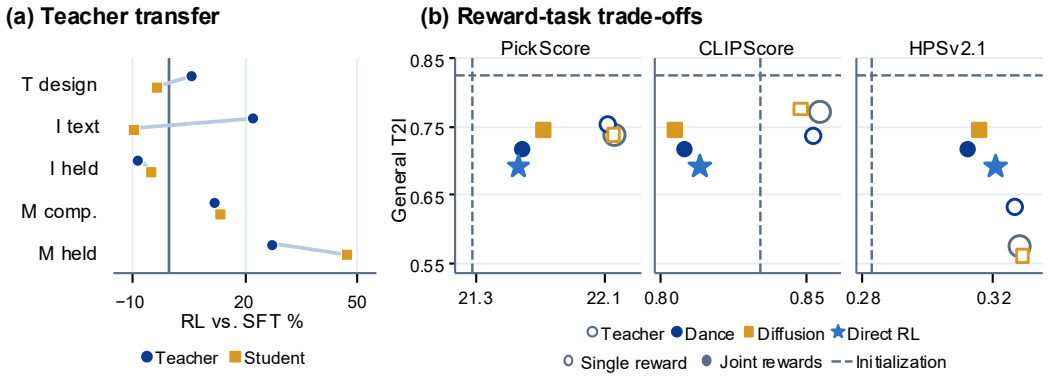

Figure 5. **Teacher-source transfer and reward to task trade-offs.** (a) RL-over-SFT changes compare teacher and student rankings across tasks. (b) Native rewards and general-T2I scores compare single-reward controls with joint OPD and direct RL.

4.3 Teacher-source transferability

Temporal supervision tests when the student is supervised; we next ask which teacher provides useful supervision. In paired DanceOPD runs, we hold the student initialization, training conditions, and optimization recipe fixed while replacing the complete SFT teacher endpoint with its RL counterpart. We test Z-Image T2I design and I2I text editing, and SD3.5-M MI2I composition. Figure 5(a) compares RL-over-SFT changes in teacher and student scores.

(i) A higher-scoring teacher can produce a weaker student. In T2I design, the RL teacher is 6.1% better than the SFT teacher on its own route, but its student is 2.9% worse than the SFT-source student. In I2I text editing, the RL teacher is 22.4% better on its own route but 7.8% worse on held-out editing; the SFT-source student is 10.0% better on text editing and 4.6% better on held-out editing than the RL-source student. Both editing students improve the trained route but regress on the held-out route relative to initialization.

(ii) The preferred teacher source reverses in MI2I composition. Unlike in design and text editing, the RL-source student is stronger in MI2I composition. The RL teacher scores 12.4% higher than the SFT teacher on composition and 27.5% higher on held-out evaluation; its student scores 13.6% and 47.5% higher than the SFT-source student on the same routes, respectively, and exceeds S_0 on both.

5 When is distillation worthwhile?

The third question is when is distillation worthwhile relative to direct joint optimization? We answer this question through two comparisons. Section 5.1 tests whether single-reward preferences survive distillation within text-to-image generation. Section 5.2 compares joint OPD, direct joint reinforcement learning, and parameter merging across text-to-image generation and image editing.

5.1 Within-interface reward transfer

We start from the same general text-to-image initialization and study PickScore, CLIPScore, and HPSv2.1 specialists (Kirstain et al., 2023; Hessel et al., 2021; Wu et al., 2023). For each reward, Figure 5(b) compares the single-reward teacher and its DanceOPD and DiffusionOPD students with the same three joint endpoints. Native rewards and the independent general text-to-image rubric use matched test inputs. The final endpoints use unequal compute, so this experiment tests reward transfer and task preservation rather than compute efficiency.

(i) Single-teacher OPD transfers reward preferences. DanceOPD and DiffusionOPD retain positive initialization-relative gains on each teacher's optimized reward. This establishes transfer of the three reward preferences but does not establish preservation of the broader task profile.

(ii) Reward recovery and independent task preservation have different orderings. Among the joint endpoints, direct reinforcement learning leads on CLIPScore and HPSv2.1, whereas DiffusionOPD leads on PickScore and general text-to-image evaluation. Nevertheless, every joint end-

	UnifiedReward 2B					CLIPScore + EditScore-4B				
	Rewards		Task scores			Rewards		Task scores		
Model	**T2I**	**I2I**	**T2I**	**I2I**	**Held**	**T2I**	**I2I**	**T2I**	**I2I**	**Held**
Student init.	3.23	15.49	0.73	6.55	0.60	0.83	4.83	0.73	6.55	0.60
T2I teacher	3.34	13.51	0.66	6.19	0.63	0.83	2.78	0.49	4.92	0.59
I2I teacher	3.19	19.81	0.75	7.01	0.87	0.81	5.73	0.71	6.91	0.87
Model merge	3.31	18.65	0.75	6.93	0.95	0.82	4.86	0.71	6.18	0.75
Direct joint RL	3.32	18.66	**0.77**	6.98	**1.16**	0.83	**6.06**	**0.73**	7.03	**1.03**
DanceOPD	3.33	18.52	0.63	7.05	0.72	0.83	5.58	0.60	6.82	0.95
Diffusion ET	3.32	18.96	0.65	7.03	0.82	0.83	5.48	0.58	**7.11**	0.74
Diffusion ED	**3.34**	**19.22**	0.64	**7.16**	0.88	**0.83**	5.54	0.61	6.91	0.87

Change vs. init. %: ≤ -20 … 0 … ≥ +20

Table 3. **Cross-interface consolidation under two reward suites.** Left: UnifiedReward-2B. Right: CLIPScore with EditScore-4B. Each block compares T2I and I2I rewards with general T2I, general I2I, and held-out I2I scores; outlines mark column winners.

point remains below the general text-to-image initialization. The HPSv2.1 specialist itself reduces general text-to-image performance by approximately 31.3% relative to initialization while increasing its optimized reward, and its students largely inherit this trade-off. Appendix F.5 provides the corresponding outputs and training curves.

5.2 Cross-interface consolidation

We next combine text-to-image generation and image editing under two reward suites: UnifiedReward and CLIPScore with EditScore-4B (Wang et al., 2025b; Wu et al., 2025). Each suite compares the common initialization, two reward specialists, their parameter midpoint (Wortsman et al., 2022; Ilharco et al., 2022), direct joint reinforcement learning, DanceOPD, and equal-time and equal-data DiffusionOPD endpoints. We report the two native rewards together with independent general T2I, general I2I, and held-out I2I scores.

(i) Direct joint reinforcement learning preserves the broadest cross-interface task profile. Under UnifiedReward, it improves general T2I, general I2I, and held-out I2I by approximately 5.5%, 6.6%, and 93.3% over initialization. Under CLIPScore with EditScore-4B, the corresponding changes are approximately 0%, 7.3%, and 71.7%. In the latter suite, EditScore increases by approximately 25.5%, while CLIPScore decreases by less than 1%.

(ii) Distillation retains narrower, metric-specific advantages. Under UnifiedReward, equal-data DiffusionOPD leads on both consolidated native rewards and general I2I. Relative to direct joint reinforcement learning, it is approximately 2.6% better on general I2I but 16.9% worse on general T2I and 24.1% worse on held-out I2I. Under the second suite, equal-data DiffusionOPD leads on consolidated CLIPScore, while time-matched DiffusionOPD leads on general I2I.

(iii) Parameter merging is sensitive to the teacher pair. The parameter midpoint improves all three independent task rubrics under UnifiedReward. Under CLIPScore with EditScore-4B, it instead reduces general T2I and general I2I while improving held-out I2I. The opposite patterns across suites do not support a general limitation of parameter merging.

Figure 19 illustrates source-preservation differences on matched editing cases, while Appendix F.5 provides text-to-image comparisons and training curves. These selected outputs do not estimate population frequencies. Existing teacher-acquisition cost can be amortized across multiple students; teachers trained solely for one student belong in the full two-stage pipeline cost.

6 Conclusion

This work provides a systematic analysis of OPD in diffusion models across thirteen routes. We find that OPD consolidates specialist gains selectively: outcomes vary with the training budget, receiving student, supervision timing, teacher source, and required task profile. Frozen-trajectory interventions show that teacher-directed corrections outperform equal-norm random perturbations, while training experiments show that fixed low-noise supervision gives the strongest overall profile and that a stronger teacher can produce a weaker student. We also identify a practical boundary between specialist reuse and broad task preservation: distillation can retain particular reward preferences, but direct joint reinforcement learning better preserves cross-interface behavior in our tested settings. We hope these findings guide future OPD design and use by the community.

AI USE STATEMENT

We used AI tools to assist with language editing and polishing of the manuscript, including improving clarity, grammar, and readability. We did not use generative AI tools for research tasks requiring disclosure under the ICLR AI policy, such as generating research data, developing the proposed method, designing or conducting experiments, or interpreting results. The authors reviewed and revised all AI-assisted text and take responsibility for the final content of this paper.

ETHICS STATEMENT

This work studies capability consolidation for text-to-image generation, single-image editing, and multi-image composition. We do not recruit human participants, conduct human-subject experiments, or collect new personal data. Training and evaluation use previously released datasets; some source datasets may contain real images, depictions of people, or user-authored prompts and editing instructions. Dataset provenance, filtering, splits, and evaluation protocols are described in Appendices B and C. The resulting models may generate inaccurate, biased, harmful, or misleading images, and image-editing capabilities could facilitate deceptive alterations of real imagery. Our experiments assess task performance and capability transfer, not safety, fairness, privacy, consent, provenance, or misuse comprehensively. Automated reward and evaluation scores are imperfect proxies and may reflect biases in the underlying models. Our results should therefore not be interpreted as evidence that these models are safe or appropriate for unrestricted deployment.

REPRODUCIBILITY STATEMENT

We describe the SD3.5-M and Z-Image configurations; specialist SFT and RL training, data provenance, and preprocessing; and the training budgets and shared controls for DanceOPD, DiffusionOPD, and FlowOPD in the main text and Appendices B.1, B.2, and B.3. Appendix C.1 specifies the 13-route test-100 benchmark, generation and judging protocols, and score aggregation. Protocols for held-out behavioral and representation diagnostics, frozen-trajectory interventions, temporal supervision, and reward evaluation are detailed in Appendices B.4, D, F.3, and B.6. We will publicly release code, checkpoints, and datasets.

REFERENCES

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

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

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

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Guanzhou Chen, Erfei Cui, Changyao Tian, Danni Yang, Ganlin Yang, Yu Qiao, Hongsheng Li, Gen Luo, and Hongjie Zhang. Scaleedit-12m: Scaling open-source image editing data generation via multi-agent framework. *arXiv preprint arXiv:2603.20644*, 2026a.

Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. *arXiv preprint arXiv:2506.18095*, 2025.

Zhekai Chen, Yuqing Wang, Manyuan Zhang, and Xihui Liu. Macro: Advancing multi-reference image generation with structured long-context data. *arXiv preprint arXiv:2603.25319*, 2026b.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4793–4801. IEEE, 2019.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

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

Siming Fu, Haojun Xu, Ruizhe He, Zheming Fu, Hualiang Wang, Jie Huang, Xiaoxiao Ma, Mingchen Zhong, Weihu Huang, Xiaoxuan He, et al. Poly-opd: Heterogeneous multi-teacher on-policy distillation for capability-selectable flow models. *arXiv preprint arXiv:2608.04349*, 2026.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Wenwang Huang, Yusen Fu, Junjie Wang, Mengfei Huang, Yulin Li, Gan Liu, Jing Cai, Yancheng He, and Zhuotao Tian. Scaling multi-reference image generation with dynamic reward optimization. *arXiv preprint arXiv:2606.26947*, 2026.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

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

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Niantong Li, Guangzheng Hu, Weixu Qiao, Ying Ba, Qichen Hong, Shijun Shen, Jinlin Wang, Fan Zhou, Jianye Kang, Xin Shang, et al. Qwen-image-bench: From generation to creation in text-to-image evaluation. *arXiv preprint arXiv:2605.28091*, 2026a.

Quanhao Li, Junqiu Yu, Kaixun Jiang, Yujie Wei, Zhen Xing, Pandeng Li, Ruihang Chu, Shiwei Zhang, Yu Liu, and Zuxuan Wu. Diffusionopd: A unified perspective of on-policy distillation in diffusion models. *arXiv preprint arXiv:2605.15055*, 2026b.

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*, 2026c.

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.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Xu Ma, Yitian Zhang, Qihua Dong, and Yun Fu. Fine-t2i: An open, large-scale, and diverse dataset for high-quality t2i fine-tuning. *arXiv preprint arXiv:2602.09439*, 2026.

Dongxing Mao, Yilin Wang, Linjie Li, Zhengyuan Yang, and Alex Jinpeng Wang. Textground4m: A prompt-aligned dataset for layout-aware text rendering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7918–7926, 2026.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5191–5198, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Yuta Oshima, Daiki Miyake, Kohsei Matsutani, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Multibanana: A challenging benchmark for multi-reference text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 448–460, 2026.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation, 2023.

Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. Pico-banana-400k: A large-scale dataset for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37226–37235, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025b.

Yuhan Wang, Siwei Yang, Bingchen Zhao, Letian Zhang, Qing Liu, Yuyin Zhou, and Cihang Xie. Gpt-image-edit-1.5 m: A million-scale, gpt-generated image dataset. *arXiv preprint arXiv:2507.21033*, 2025c.

Xinyu Wei, Kangrui Cen, Hongyang Wei, Zhen Guo, Bairui Li, Zeqing Wang, Jinrui Zhang, and Lei Zhang. Mico-150k: A comprehensive dataset advancing multi-image composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29695–29706, 2026.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998. Pmlr, 2022.

Keming Wu, Sicong Jiang, Max Ku, Ping Nie, Minghao Liu, and Wenhu Chen. Editreward: A human-aligned reward model for instruction-guided image editing. *arXiv preprint arXiv:2509.26346*, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025.

Keming Ye, Zhipeng Huang, Canmiao Fu, Qingyang Liu, Jiani Cai, Zheqi Lv, Chen Li, Jing Lyu, Zhou Zhao, and Shengyu Zhang. Unicedit-10m: A dataset and benchmark breaking the scale-quality barrier via unified verification for reasoning-enriched edits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37279–37289, 2026.

Hui Zhang, Juntao Liu, Zongkai Liu, Liqiang Niu, Fandong Meng, Zuxuan Wu, and Yu-Gang Jiang. Weedit: A dataset, benchmark and glyph-guided framework for text-centric image editing. In *European Conference on Computer Vision*, pp. 488–507. Springer, 2026.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, volume 2026, pp. 134129–134150, 2026.

Wei Zhou, Xiongwei Zhu, Lingdong Kong, Bo Chen, Lei Zhang, Yongyuan Liang, Xiaoxia Hou, Ye Tian, Xian Sun, Yingshuo Wang, et al. On-policy self-distillation in diffusion models. *arXiv preprint arXiv:2608.24646*, 2026a.

Wei Zhou, Xiongwei Zhu, Zelin Xu, Bo Dong, Lixue Gong, Yongyuan Liang, Meng Chu, Leigang Qu, Lingdong Kong, Wei Liu, et al. Danceopd: On-policy generative field distillation. *arXiv preprint arXiv:2606.27377*, 2026b.

A Related work

On-policy distillation in diffusion models. On-policy distillation (OPD) is an emerging knowledge distillation paradigm in which teachers supervise a student at states visited along the student's own denoising trajectories. In diffusion models, it offers a way to consolidate complementary generation and editing capabilities from task-specific teachers into a shared student. FlowOPD combines routed trajectory supervision with manifold regularization; DiffusionOPD matches teacher and student transition means; and DanceOPD uses sparse velocity supervision across generation and editing (Fang et al., 2026; Li et al., 2026b; Zhou et al., 2026b). Other on-policy diffusion work uses self-distillation to continue tuning few-step models (Jiang et al., 2026), converts reward feedback into on-policy targets (Zhou et al., 2026a), and consolidates heterogeneous teachers across latent spaces (Fu et al., 2026). Together, these studies do not establish which teacher gains transfer to trained and held-out routes, how supervision timing and teacher source affect transfer, or when OPD is preferable to direct joint reinforcement learning. We study these questions through budget-controlled comparisons, held-out behavioral evaluation, frozen-trajectory interventions, temporal-supervision and teacher-source experiments, and comparisons with direct joint reinforcement learning.

Capability transfer in distillation. How teacher capabilities transfer to student behavior has long been studied in knowledge distillation, from model compression to language model OPD (Hinton et al., 2015; Agarwal et al., 2024). Prior studies show that student capacity and target learnability can constrain transfer: intermediate hints and teacher assistants can bridge large capacity gaps, while scaling analyses and long reasoning traces reveal further limits (Cho & Hariharan, 2019; Mirzadeh et al., 2020; Busbridge et al., 2025; Li et al., 2025). In large language model OPD, Li et al. (2026c) further relate transfer to teacher and student compatibility and capabilities beyond the student's existing behavior. These findings do not determine which teacher behaviors diffusion OPD transfers at student-visited states, whether gains persist on held-out routes, or how supervision timing and teacher source affect transfer. We therefore compare student behavior with its exact initialization and routed teachers, then test how teacher-directed corrections, query timing, and teacher source change task outcomes.

B Training and evaluation protocols

B.1 Supervised fine-tuning

Backbones and optimization. We train SFT specialists from SD3.5-M (Esser et al., 2024) and the original Z-Image (Cai et al., 2025). All runs use 512×512 images, $8\times$A100 GPUs, global batch size 32, rank-32 LoRA, and AdamW with learning rate 10^{-4}, weight decay 10^{-4}, gradient clipping at 1.0, and 100 warm-up steps followed by a constant schedule. The variational autoencoder and text encoders are frozen. Z-Image uses uniform flow timesteps, its native text encoder with length 512, and attention q/k/v/out LoRA. SD3.5-M uses logit-normal timesteps, CLIP-L, CLIP-G, and T5 with length 256, joint/context attention LoRA, and learned edit-condition embeddings. Evaluation uses 50 steps with guidance 4.0 for Z-Image and 40 steps with guidance 4.5 for SD3.5-M.

T2I data and specialists. The five routes are general generation, text rendering, portrait, design, and realism. Fine-T2I supplies the general, portrait, design, and realism candidates (Ma et al., 2026); TextGround4M supplies text-rendering candidates (Mao et al., 2026). After filtering and de-duplication, each route contains $100,000$ examples. Each specialist starts from its public backbone and trains for one epoch, or $3,125$ updates.

I2I data and specialists. The general pool has $1,949,841$ source, instruction, and target pairs from GPT-Image-Edit, Pico-Banana, Nano-consistent-150k, ShareGPT-4o-Image, ScaleEdit-12M, and UnicEdit-10M (Wang et al., 2025c; Qian et al., 2026; Chen et al., 2025; 2026a; Ye et al., 2026). A parent first trains for $9,375$ updates on 300K balanced local, global, and text edits, then for one epoch on the full pool, ending at $60,932$ updates. Native labels are retained; InternVL3.5 labels only the $329,683$ unlabeled pairs (Wang et al., 2025a). After de-duplication, local, global, and text specialists warm-start from the parent and end at $36,736$, $14,230$, and $24,674$ updates. Only the target latent is noised and included in the loss; the source remains a clean condition.

MI2I data and specialists. The $509,902$-example general mixture draws from MACRO, Echo-4o, and MICo-150K (Chen et al., 2026b; Ye et al., 2025; Wei et al., 2026). It starts from the general I2I endpoint and trains for one epoch, ending at $15,934$ updates. Specialist routing also includes OmniRef (Huang et al., 2026) and assigns each example once: pose, viewpoint, and layout requests go to spatial; multi-subject identity and attribute binding to identity; and object, style, background, lighting, and interaction fusion to composition. The three specialists end at $7,742$, $6,054$, and $4,816$ updates. Up to ten ordered references use clean condition tokens with explicit roles and a shared 512^2 condition-pixel budget; noise and loss remain confined to the target.

B.2 Reinforcement learning

We use DiffusionOPSD (Zhou et al., 2026a), initialized from the corresponding SFT endpoint and trained separately for each route and backbone. The frozen reward model is InternVL3.5-38B (Wang et al., 2025a), served independently from six-GPU policy workers. Each update samples 100 train-only conditions and $N_{\text{cand}} = 6$ candidate images per condition at 512×512 resolution, using ten sampler steps and guidance scale 1. SD3.5-M uses deterministic DPM2 and Z-Image its native flow solver. The budget is at most 20 updates, with AdamW learning rate 3×10^{-4}, weight decay 10^{-4}, and gradient clipping at 1.0.

The reward model receives the instruction, source image when applicable, ordered references with their roles, and one candidate image. It is frozen in evaluation mode and returns a holistic task score on the integer scale $\{1, 2, 3, 4, 5\}$. We read the score from the last-position logits rather than from an unconstrained generated explanation. For score $s \in \{1, \dots, 5\}$, let λ_s be the logit of its corresponding token and $\pi_s = \frac{\exp(\lambda_s)}{\sum_{h=1}^{5} \exp(\lambda_h)}$ its probability normalized over the five score tokens. The expected scalar reward is $\widehat{R}_{\text{exp}} = \sum_{s=1}^{5} s\,\pi_s$. The differentiable reward path uses $\widehat{R}_{\text{exp}}$; the JSON readout instead returns the selected score $\widehat{R}_{\text{argmax}} = \arg\max_{s \in \{1,\dots,5\}} \lambda_s$. The scorer also verifies sufficient probability mass on the five score tokens and rejects non-finite or malformed responses. The rubric, model revision, score-token protocol, and malformed-response handling are fixed across runs.

We monitor 12 fixed conditions disjoint from RL training after every update, with minimum update of 3, an improvement threshold 0.01, and patience 3. After the reward no longer increases, we select the best checkpoint observed after the plateau, i.e., the saved checkpoint with the highest monitor expected reward. This selected checkpoint is exported as the RL endpoint used in Table 4; benchmark test and held-out examples are never used for reward computation or endpoint selection.

B.3 Common on-policy distillation experimental design

Backbones and consolidation mixtures. TI combines `T2I-GENERAL` and `I2I-GENERAL`, with I2I balanced between local and global edits; TIM adds `MI2I-COMPOSITION`. Each route contributes $3,000$ conditions, yielding 188 TI and 282 TIM updates. TI starts from the `I2I-GLOBAL` RL bundle; TIM starts from an independently evaluated equal-weight merge of the `I2I-GLOBAL` and `MI2I-COMPOSITION` RL bundles. Initialization is identical within each mixture, but differs between mixtures along with route composition.

Shared comparison controls. DanceOPD, DiffusionOPD, and FlowOPD share route data, rank-32 student LoRA with $\alpha = 64$, optimizer, and rollout length. During distillation, student and routed-teacher guidance are both 1.0, aligning adapter capacity and guidance strength. This differs from some SD3.5-M recipes and the original DanceOPD adapter size; we do not use asymmetric guidance absorption. The methods retain their native sparse velocity, dense transition-mean, and stochastic transition objectives, respectively, including their weighting and regularization.

B.4 Capability-absorption diagnostic protocol

Behavioral audit. We pair generations by sample identifier, condition, source and reference order, sampler, and endpoint under the original rubrics. The same-image statistic averages each image's lowest criterion score over 100 examples per held-out suite, without a new judge or threshold.

CKA. For each backbone and mixture, we use the first 32 registered I2I-GENERAL and I2I-HELDOUT conditions. We cache initialization trajectories, capture steps 1, 5, and 9 of the ten-step sampler, and mean-pool target tokens at six approximately uniform transformer depths. Each of the 40 endpoint comparisons yields 18 linear-CKA values.

DINOv2 correspondence. We freeze `dinov2_vitb14_reg`, resize and pad inputs to 518 pixels with official normalization, and exclude padding and register tokens. Mutual-nearest-neighbor patch matches compare outputs with references and teachers; at most 12 appear per diagnostic panel. Color represents frozen visual-feature similarity, not attention or causal attribution. The supplemental examples in Figure 9 reuse existing generations and rubric judgments.

B.5 Teacher performance

Table 4 reports frozen-teacher point scores on the 13-task benchmark. All supported route cells are evaluated; text-to-image-only checkpoints lack an edit interface and are marked n/a.

Reward-defined teachers and direct RL. Within T2I, PickScore, CLIPScore, and HPSv2.1 teachers each use 10K General-T2I conditions (Kirstain et al., 2023; Hessel et al., 2021; Wu et al., 2023). Single-teacher OPD uses 3K conditions per reward; joint OPD uses 1K per teacher; direct multi-reward RL uses the 10K pool with equal weights after frozen train-only normalization. Cross-interface comparisons use UR2B, pairing UnifiedReward-2.0-2B with UnifiedReward-Edit-2B, and EC4, pairing CLIPScore with EditScore-4B (Wang et al., 2025b; Wu et al., 2025). Each route-specific teacher uses 10K conditions and 100 updates. Cross-interface OPD uses 3K conditions per route and 188 updates; direct joint RL alternates T2I and I2I for 100 total updates, consuming 5K conditions per route with route-specific advantage moments. The model-merge control averages teachers' effective LoRA updates with coefficients 0.5 and 0.5. We report integration and full teacher-acquisition costs; only the two OPD equal-time endpoints are wall-clock matched.

Endpoint labels in reward comparisons. In Table 3, ET is DiffusionOPD at step 38, matched to DanceOPD wall clock, and ED is its step-188 equal-data endpoint. Both start from the archived General-I2I RL initialization. For I2I evaluation of a T2I reward teacher, its T2I LoRA is paired with the unchanged inherited edit adapter, without additional training. Scores come from independent full13 evaluation and matched raw test-100 rescoring, rounded to two decimals.

B.6 Reward-evaluation data contracts

Reward evaluation. Reward-model scores use fixed test inputs separate from the benchmark task scores in Appendix C. T2I reward evaluation uses the General-T2I test-100 split from Qwen-Image-Bench (Li et al., 2026a). For I2I, UnifiedReward-Edit-2B and EditScore-4B use a balanced General-I2I test-100 manifest with 50 local and 50 global examples selected by deterministic stratification from the frozen GEdit-Bench splits, released as `manifests/i2i_general_test_100.jsonl`. This composite evaluates both local and scene-level edits rather than the global subset alone.

Each edit reward model receives the instruction, source image, and matched frozen generation, but not the target image. We macro-average the frozen score-token readout over the manifest. Test examples are excluded from training, reward computation, normalization, and checkpoint selection. The composite changes only reward evaluation, not training mixtures or benchmark task-score splits.

C Evaluation

C.1 Benchmark and shared evaluation protocol

The benchmark contains five text-to-image, four image-editing, and four multi-image routes. On each route supported by its interface, every checkpoint generates one image for each example in a fixed test-100 split. Unsupported interfaces are marked as not applicable. Sample identifiers, input order, generation parameters, judge prompts, and aggregation are shared by supervised fine-tuning and reinforcement learning. We first average within the benchmark's native facet, operation,

Table 4. **Specialist performance across 13 routes.** Paired SFT and RL endpoints are evaluated on five T2I, four I2I, and four MI2I routes. Cell colors show RL changes from SFT; gray denotes references or unsupported interfaces.

		T2I					I2I				MI2I			
	Stage	General	Text	Portrait	Design	Realism	Local	Global	Text	Held-out	Identity	Comp.	Spatial	Held-out
SD3.5-M														
	Base	0.73	0.35	0.66	0.79	1.06	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
General I2I	SFT	0.74	0.18	0.62	0.52	0.94	5.27	6.07	2.12	2.67	3.65	3.91	1.54	3.49
General MI2I	SFT	0.82	0.18	0.61	0.60	0.97	3.92	4.36	1.39	2.69	5.75	5.30	4.39	3.39
T2I-GENERAL	SFT	0.85	0.39	0.65	0.85	1.00	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	RL	0.83	0.41	0.67	0.84	0.99	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
T2I-TEXT	SFT	0.76	0.41	0.64	0.68	0.95	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	RL	0.73	0.43	0.57	0.67	0.85	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
T2I-PORTRAIT	SFT	0.80	0.44	0.69	0.81	0.92	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	RL	0.79	0.46	0.64	0.74	0.94	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
T2I-DESIGN	SFT	0.79	0.32	0.69	0.83	0.97	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	RL	0.83	0.39	0.64	0.77	0.99	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
T2I-REALISM	SFT	0.80	0.21	0.68	0.69	1.01	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	RL	0.77	0.29	0.65	0.79	0.92	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
I2I-LOCAL	SFT	0.77	0.23	0.67	0.58	0.91	5.47	6.08	2.17	2.78	3.84	4.20	1.51	3.41
	RL	0.80	0.22	0.61	0.70	0.87	5.80	6.54	0.74	0.42	4.21	4.52	1.42	3.34
I2I-GLOBAL	SFT	0.74	0.32	0.57	0.66	0.89	5.21	6.45	2.29	3.33	4.08	4.22	1.17	3.09
	RL	0.73	0.18	0.57	0.69	0.90	5.17	6.55	0.63	0.60	4.07	4.55	1.23	3.20
I2I-TEXT	SFT	0.46	0.09	0.49	0.36	0.53	4.32	5.70	3.58	4.28	2.91	2.75	1.32	2.17
	RL	0.60	0.14	0.48	0.49	0.60	5.11	6.38	1.08	1.09	3.19	3.05	1.02	2.05
MI2I-IDENTITY	SFT	0.76	0.28	0.69	0.64	0.94	4.15	4.93	1.73	3.27	6.10	5.66	4.63	4.28
	RL	0.83	0.28	0.68	0.69	0.87	4.49	6.00	0.95	0.98	6.42	5.92	4.44	4.64
MI2I-COMP.	SFT	0.76	0.21	0.60	0.59	0.87	4.17	4.65	1.53	3.01	5.77	5.00	3.93	3.38
	RL	0.87	0.32	0.65	0.67	0.88	4.93	5.68	0.84	1.09	5.78	5.62	3.97	4.31
MI2I-SPATIAL	SFT	0.80	0.27	0.58	0.65	0.92	4.29	4.82	1.63	3.02	5.68	5.28	4.24	3.77
	RL	0.80	0.25	0.57	0.62	0.92	4.85	5.95	0.88	0.89	5.84	5.18	4.72	3.89
Z-Image														
	Base	0.99	0.72	0.85	0.86	1.00	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
General I2I	SFT	0.93	0.70	0.82	0.79	1.00	5.94	6.90	3.35	5.05	5.17	4.98	2.34	4.71
General MI2I	SFT	1.01	0.82	0.92	0.87	1.08	6.15	7.17	3.42	5.42	6.72	6.23	5.70	4.92
T2I-GENERAL	SFT	0.99	0.96	0.92	1.01	1.11	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	RL	1.02	0.97	0.91	1.06	1.09	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
T2I-TEXT	SFT	0.97	0.65	0.88	0.89	1.02	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	RL	0.95	0.69	0.84	0.91	1.01	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
T2I-PORTRAIT	SFT	0.96	0.99	0.94	1.02	1.01	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	RL	1.02	1.01	0.94	1.02	1.05	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
T2I-DESIGN	SFT	1.00	0.97	0.91	0.98	1.08	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	RL	1.02	1.03	0.93	1.04	1.10	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
T2I-REALISM	SFT	0.91	0.75	0.88	0.84	1.10	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	RL	0.91	0.79	0.86	0.75	1.12	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
I2I-LOCAL	SFT	0.98	0.69	0.85	0.83	0.98	5.83	6.72	3.33	5.05	5.20	5.02	2.43	4.82
	RL	0.99	0.76	0.84	0.86	1.01	6.31	7.16	3.49	4.14	5.11	5.02	2.49	4.95
I2I-GLOBAL	SFT	0.94	0.80	0.77	0.87	0.99	6.11	7.21	3.90	5.47	5.20	4.89	2.45	4.63
	RL	0.96	0.80	0.81	0.86	0.97	6.75	7.84	4.43	4.14	5.31	4.96	2.46	4.46
I2I-TEXT	SFT	0.95	0.64	0.80	0.70	0.96	6.19	6.49	4.00	4.49	4.95	4.50	1.75	4.05
	RL	0.97	0.66	0.81	0.69	1.00	6.83	7.14	5.14	4.09	4.97	4.58	1.74	4.09
MI2I-IDENTITY	SFT	1.03	0.95	0.88	0.88	1.07	5.85	7.14	3.27	5.08	7.08	6.34	5.70	5.38
	RL	1.02	0.89	0.87	0.91	1.02	6.52	7.73	3.15	4.43	7.17	6.45	5.64	5.45
MI2I-COMP.	SFT	0.93	0.91	0.89	0.81	1.07	6.00	7.15	3.28	5.48	6.92	6.17	5.82	5.23
	RL	0.93	0.96	0.87	0.80	1.03	6.67	7.69	3.02	4.50	6.84	6.17	5.74	5.22
MI2I-SPATIAL	SFT	1.02	0.81	0.90	0.85	1.06	6.21	7.16	3.41	5.19	6.89	6.21	5.82	5.20
	RL	0.98	0.81	0.86	0.85	1.07	6.76	7.76	3.45	4.49	6.84	6.12	5.87	5.01

RL change vs. paired SFT (%) ≤ -20 0 ≥ +20

or reference-count strata and then macro-average the strata. The image-editing and multi-image held-out splits contain no training or reinforcement-learning reward examples.

Text-to-image judging. We sample 100 prompts per route from Qwen-Image-Bench (Li et al., 2026a). The samples cover five balanced facets each for general generation, portrait, and realism; approximately balanced design facets; and 34%, 33%, and 33% text-rendering prompts across three non-font subsets. Q-Judger (Li et al., 2026a) evaluates only the route's registered facets with deterministic decoding at temperature 0, restricted to the highest-probability token. Optical character recognition and structural measurements are diagnostics and do not change the reported route score.

Image-editing judging. Local and global editing use 100 real-user examples each from GEdit-Bench (Liu et al., 2025). Local editing is stratified over subject addition, removal, replacement, color and material changes, motion changes, and human editing; global editing uses balanced background, style, and tone-transfer subsets. Text editing uses 100 WeEdit examples (Zhang et al., 2026), split equally between bilingual and multilingual cases and stratified over eight operations. Our I2I held-out set has 100 examples across structure, viewpoint, and process reasoning families. GPT-5.5 receives the instruction, source, and candidate in a fixed order and returns task success, target correctness, source/non-target preservation, and visual-quality axes; text editing also scores requested content and legibility.

Multi-image judging. Identity and spatial editing use 100 examples each from the customization and spatial evaluation sets of MACRO (Chen et al., 2026b), stratified by reference count. Composition uses 100 examples from MultiBanana (Oshima et al., 2026), stratified by task, reference count, and domain. Our MI2I held-out set contains 100 irreducible multi-reference cases across four families. GPT-5.5 receives the ordered references, role labels, instruction, and candidate, and returns per-reference fidelity, identity/role binding, relation/layout correctness, omission/confusion, and visual-quality axes. Malformed outputs are retried under the same frozen prompt; valid axes are aggregated by the route-specific rubric.

D TRAJECTORY INTERVENTION RESULTS

Table 5 gives absolute scores for three preselected large-gap failures used for illustration. The aggregate diagnostic uses randomly picked 480 cases: 180 T2I, 180 I2I, and 120 MI2I. For each case, task gain is the branch's native-rubric score minus the unmodified student's score. Terminal correction is the reduction in Euclidean distance to the full teacher's final target latent; positive values indicate movement toward the teacher. Figure 10 shows case-level gains, while Figure 3 separates task gain from terminal correction by task family.

Table 5. **Single step interventions on three selected failures.** Columns compare the full teacher, unmodified student, and teacher steps applied at early, middle, and late denoising stages.

Change vs. student (%) ≤ -20 0 ≥ +20

	Reference endpoints		Teacher-step timing		
Case	Teacher	Student	Early	Mid	Late
I2I-TEXT	8.25	1.25	2.00	1.75	1.75
MI2I-SPATIAL-1	8.00	1.60	2.40	2.80	2.40
MI2I-SPATIAL-2	7.60	1.20	2.10	1.80	2.10

E DETAILED RESULTS

Table 6. **Held-out multi-criterion performance.** I2I and MI2I entries report mean per-image minimum rubric scores over 100 conditions per suite. Outlined cells mark the best student.

Best student per suite · Equal time · Equal data

Backbone	Mixture	Budget	I2I held-out				MI2I held-out			
			Dance	Flow	Diffusion	Teacher	Dance	Flow	Diffusion	Teacher
SD3.5-M	T2I + I2I	⏱	0.13	**0.27**	0.18	0.19	1.59	**1.70**	1.61	n/a
		🗄	0.13	**0.26**	0.19	0.19	1.59	1.59	**1.69**	n/a
SD3.5-M	T2I + I2I + MI2I	⏱	0.22	**0.32**	0.20	0.19	2.40	**2.65**	2.41	2.02
		🗄	0.22	**0.34**	0.18	0.19	**2.40**	2.28	2.13	2.02
Z-Image	T2I + I2I	⏱	**2.29**	2.13	2.18	2.20	2.64	2.57	**2.67**	n/a
		🗄	2.29	2.27	**2.32**	2.20	2.64	2.67	**2.73**	n/a
Z-Image	T2I + I2I + MI2I	⏱	2.25	2.43	**2.54**	2.20	2.61	**2.77**	2.74	2.66
		🗄	2.25	2.45	**2.49**	2.20	2.61	**2.81**	2.64	2.66

Table 7. **Temporal supervision across** 13 **routes.** The upper block compares fixed query schedules, and the lower block compares query-matched uniform and disagreement weighting.

Change vs. block uniform (%) · ≤ -20 0 ≥ +20 · 0.88 Best per block

Run / schedule	T2I General	T2I Text	T2I Portrait	T2I Design	T2I Realism	I2I Local	I2I Global	I2I Text	I2I Held-out	MI2I Identity	MI2I Comp.	MI2I Spatial	MI2I Held-out
Fixed query-time distributions													
Low-time	0.78	**0.32**	0.56	**0.68**	**0.92**	**5.51**	**6.84**	**1.00**	**0.93**	4.07	**4.70**	**1.55**	**4.28**
Middle-time	0.78	0.21	**0.61**	0.60	0.89	5.37	6.33	0.91	0.82	4.15	4.40	1.25	4.01
High-time	**0.78**	0.28	0.57	0.66	0.87	5.39	6.54	0.93	0.90	**4.22**	4.65	1.47	4.08
Uniform	0.77	0.26	0.56	0.63	0.90	5.32	6.55	0.92	0.75	4.09	4.56	1.40	3.81
Query-matched weighting rules													
Uniform-weight	**0.76**	0.26	0.64	**0.66**	**0.91**	5.21	**6.08**	0.88	0.87	**4.18**	4.42	1.79	3.97
Disagreement-weight	0.72	**0.26**	**0.64**	0.62	0.86	**5.25**	6.06	**0.91**	**0.96**	4.17	**4.67**	**1.84**	**4.08**

Table 8. **Teacher-source transfer across three tasks.** Each block compares SFT and RL teachers with their corresponding OPD students.

Change vs. init. (%) · ≤ -20 0 ≥ +20 · 0.88 Best OPD: own / held-out

Setting / backbone	Role	T2I General	T2I Text	T2I Portrait	T2I Design	T2I Realism	I2I Local	I2I Global	I2I Text	I2I Held-out	MI2I Identity	MI2I Comp.	MI2I Spatial	MI2I Held-out
T2I-DESIGN Z-Image	Init	0.99	0.72	0.85	0.86	1.01	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	SFT teacher	1.00	0.97	0.91	0.98	1.08	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	RL teacher	1.02	1.03	0.93	1.04	1.10	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	SFT → OPD	1.03	0.99	0.90	**1.04**	1.11	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	RL → OPD	1.02	0.84	0.92	1.01	1.13	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
I2I-TEXT Z-Image	Init	0.93	0.70	0.82	0.79	1.00	5.94	6.90	3.35	5.05	5.17	4.98	2.34	4.71
	SFT teacher	0.95	0.64	0.80	0.70	0.96	6.19	6.49	4.01	4.49	4.95	4.50	1.75	4.05
	RL teacher	0.99	0.67	0.78	0.65	0.95	6.82	7.14	4.91	4.14	4.99	4.63	1.81	4.06
	SFT → OPD	0.97	0.83	0.84	0.80	1.01	6.98	7.60	**5.05**	**3.38**	4.67	4.45	1.72	4.09
	RL → OPD	1.00	0.70	0.80	0.74	1.01	6.96	7.60	4.59	3.23	4.75	4.55	1.71	4.01
MI2I-COMP. SD3.5-M	Init	0.82	0.18	0.61	0.60	0.97	3.92	4.36	1.39	2.69	5.75	5.30	4.39	3.39
	SFT teacher	0.76	0.21	0.60	0.59	0.87	4.17	4.65	1.53	3.01	5.77	5.00	3.93	3.38
	RL teacher	0.87	0.32	0.65	0.67	0.88	4.93	5.68	0.84	1.09	5.78	5.62	3.97	4.31
	SFT → OPD	0.78	0.21	0.66	0.59	0.98	3.80	3.85	0.50	0.40	5.53	4.84	4.27	3.05
	RL → OPD	0.79	0.33	0.64	0.69	0.97	4.95	5.80	0.81	1.11	5.66	**5.50**	4.15	**4.50**

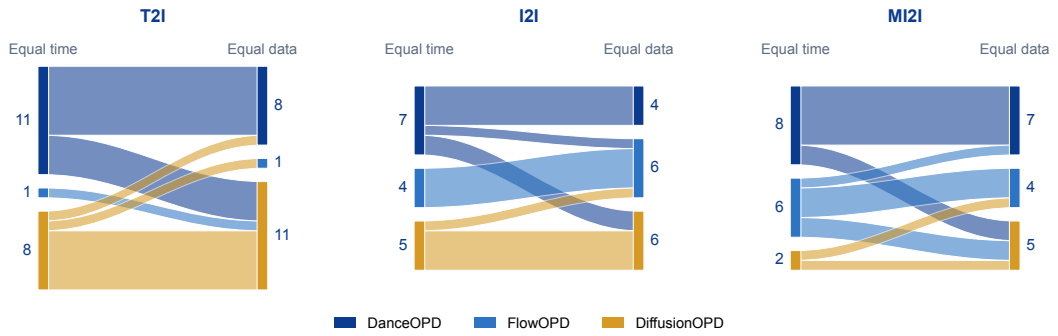

Figure 6. **Budget-dependent winner transitions.** The T2I, I2I, and MI2I panels track method winners from equal time to equal data across 20, 16, and 16 comparisons, respectively.

Table 9. **Within-T2I reward transfer and joint optimization.** Rows compare initialization, reward specialists, single- and joint-teacher OPD, and direct multi-reward RL on PickScore, CLIPScore, HPSv2.1, and general-T2I performance.

Change vs. init. (%) ≤ -20 0 ≥ +20

Model	PickScore	CLIPScore	HPSv2.1	General-T2I
Student initialization	21.28	0.83	0.28	0.83
Single-reward acquisition and transfer				
PickScore teacher	22.17	0.80	0.31	0.74
PickScore → DanceOPD	22.13	0.80	0.30	0.75
PickScore → DiffusionOPD	22.17	0.80	0.31	0.74
CLIPScore teacher	21.08	0.85	0.28	0.77
CLIPScore → DanceOPD	21.11	0.85	0.28	0.74
CLIPScore → DiffusionOPD	20.96	0.85	0.28	0.77
HPSv2.1 teacher	21.16	0.76	0.33	0.57
HPSv2.1 → DanceOPD	21.37	0.78	0.33	0.63
HPSv2.1 → DiffusionOPD	21.25	0.76	0.33	0.56
Joint multi-reward comparison				
Joint teachers → DanceOPD	21.59	0.81	0.31	0.72
Joint teachers → DiffusionOPD	21.73	0.80	0.32	0.75
Direct multi-reward RL	21.57	0.81	0.32	0.69

F More results

F.1 Consolidation across budgets

Figure 6 tracks method winners between matched equal-time and equal-data comparisons; ribbon color denotes the equal-time winner and width denotes route count. Figure 7 shows nine selected teacher-gap cases illustrating partial transfer, not its prevalence.

Figure 1A normalizes Z-Image route scores to initialization; IH and MH denote held-out suites. Panel B reports image-editing gains over eighteen selected intervention cases. Panel C uses the UR2B cross-interface I2I columns of Table 3: each endpoint's reward and independent held-out task score are divided by the corresponding score of the exact initialization. The reward/held-out ratios are $(1.1956, 1.2000)$ for DanceOPD, $(1.2408, 1.4667)$ for equal-data DiffusionOPD, and $(1.2046, 1.9333)$ for direct joint RL. The endpoints use different budgets and no cross-metric average; the axis begins at 0.5 only for display.

F.2 Representation and task-constraint examples

Figure 8a reports twenty endpoint-suite comparisons per mixture at sampler steps 1, 5, and 9; relative depth aligns the two backbones without equating layer indices. Figure 9 applies the frozen DINOv2 protocol to three cases from Figure 7 and one held-out courier-robot case. Each source or reference

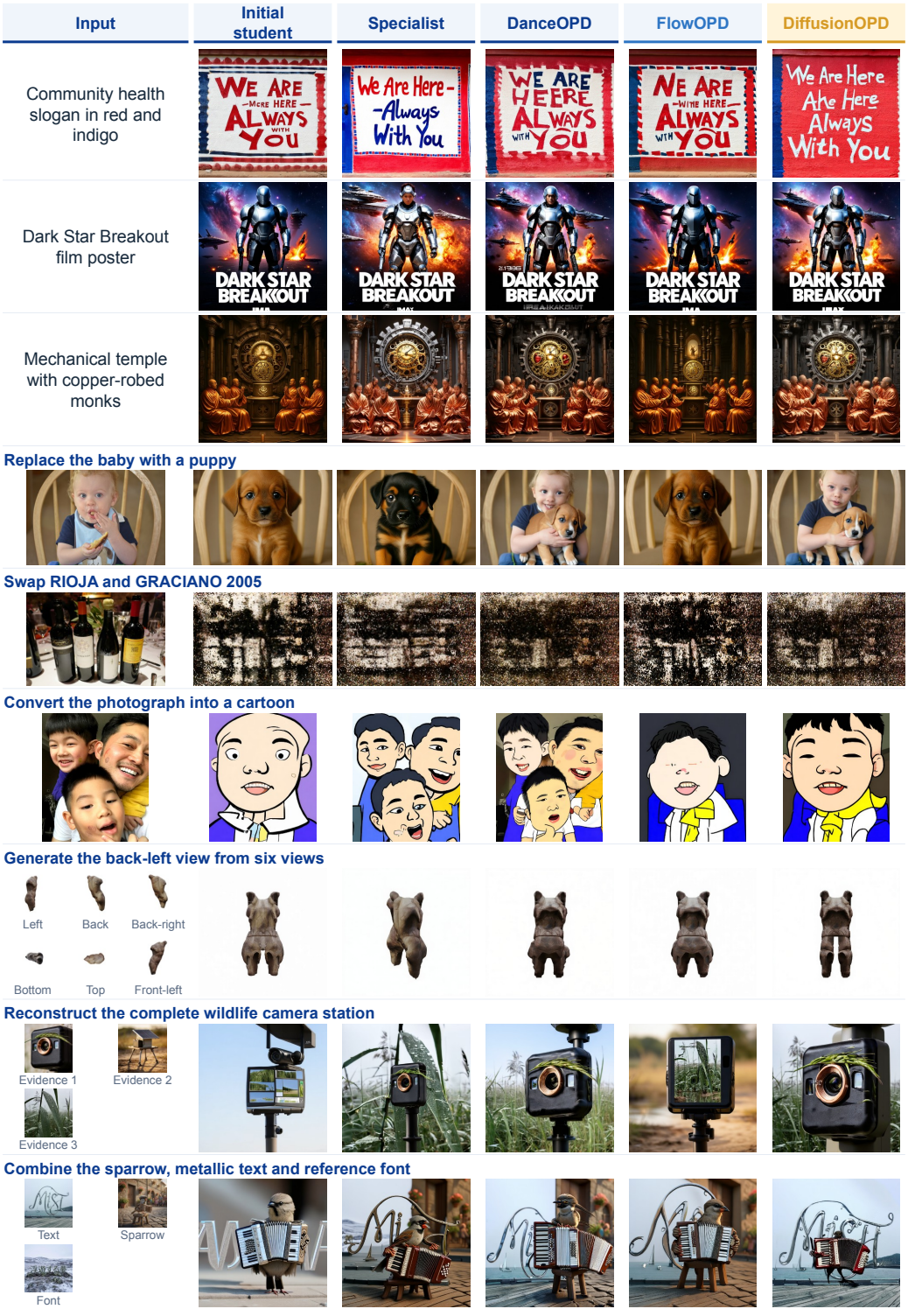

Figure 7. **Qualitative comparison of capability consolidation.** Nine SD3.5-M TIM cases compare initialization, specialists, and equal-data OPD students.

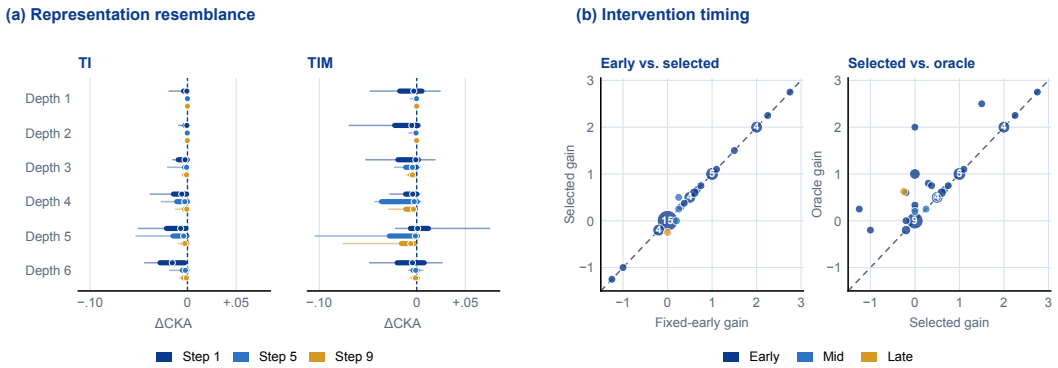

Figure 8. **Representation resemblance and intervention timing.** (a) Median ΔCKA across sampler states and depths, with interquartile and 5th–95th-percentile ranges. (b) Gains from fixed-early, maximum-disagreement, and oracle intervention timing on native rubrics.

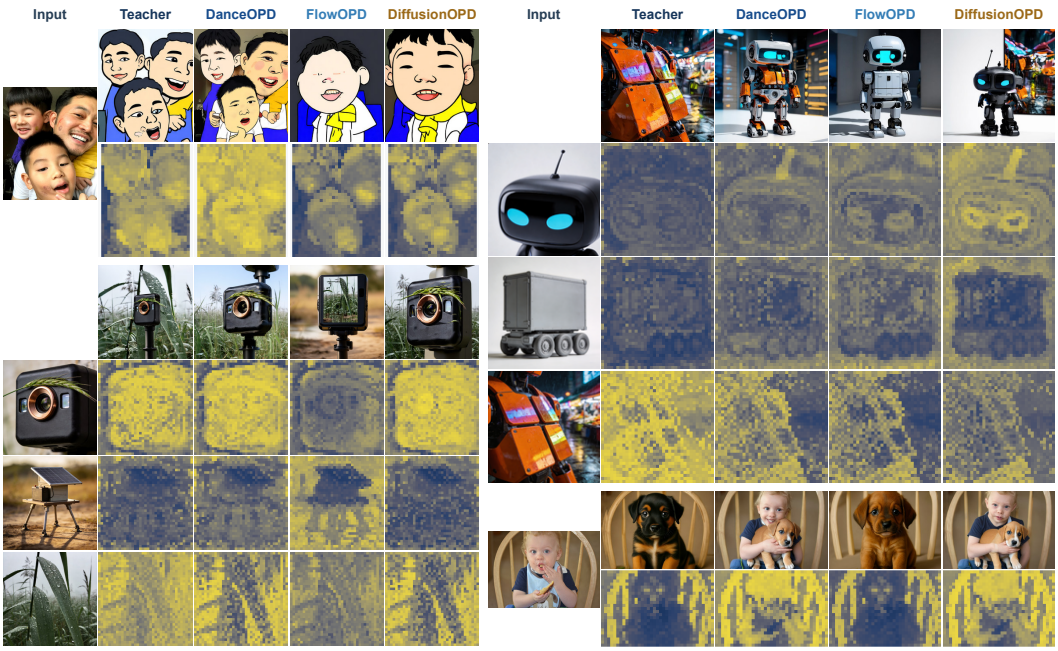

Figure 9. **DINO correspondence and task constraints.** Four matched editing and multi-image cases compare teacher and OPD outputs with reference-patch best-match maps; blue-to-yellow values indicate cosine similarity on a shared $[0, 1]$ scale, not task success.

patch is colored by its best output-patch cosine similarity on a shared $[0, 1]$ scale, excluding padding. Maps retain reference coordinates and may reuse output patches; they measure visual resemblance, not instruction success.

In the cartoon-editing case, student-to-teacher matches range from 0.742 to 0.769 while task scores range from 3.00 to 8.25; similar resemblance accompanies different subject preservation. In the baby-to-puppy replacement, DanceOPD and DiffusionOPD retain the baby while adding a puppy. Their reference matches, 0.711 and 0.689, exceed the teacher's 0.471, but their task scores are 4.25 versus the teacher's 8.50. In the multi-reference courier-robot case, DiffusionOPD matches the face reference more strongly than the teacher (0.601 versus 0.358), yet all three students generate legs instead of the requested wheels. These selected failures illustrate why correspondence cannot substitute for held-out task judgments.



Figure 10. **Intervention gains by 3 different task families.** Rows separate T2I, I2I, and MI2I tasks; columns show early, middle, and late denoising intervention. Dots denote cases, boxes interquartile ranges, diamonds means, and whiskers Tukey ranges.

F.3 Trajectory interventions and temporal training

Figure 8b compares fixed-early, maximum-disagreement, and oracle one-step timing on the same cases. The oracle selects the best measured intervention and is not a deployable rule. Across all cases, teacher-directed corrections exceed equal-norm random controls by 0.269, 0.125, and 0.076 on the respective native task scales. Fixed-early and disagreement-selected means tie in T2I and I2I; in MI2I, fixed early scores 0.297 versus 0.281 for selected timing. Figure 10 shows all 4320 scored branches by task family and timing; Figure 3 reports their mean task gains and terminal corrections.

Figure 11 shows six matched image-editing cases across both backbones and all three objectives. Random and one-step controls use the early position; three-step windows vary from early to late.

Temporal-supervision runs share the SD3.5-M image-editing RL initialization and a frozen multi-image composition RL teacher. They use the same optimizer, $3,000$ General-I2I conditions, and 94 updates. Fixed schedules sample denoising progress uniformly or from $\mathrm{Beta}(5,2)$, $\mathrm{Beta}(5,5)$, and $\mathrm{Beta}(2,5)$ for low-, middle-, and high-noise states, respectively. Larger progress means lower noise. The weighting comparison instead fixes queries at 20%, 50%, and 80% progress and changes only the detached weights in Section 4.2. Figure 12 shows nine-update rolling medians; losses with different targets are not directly comparable.

F.4 Teacher-source optimization and transfer

Figure 13 juxtaposes teacher-specific fitting traces and downstream student-score changes. Figure 5 includes an editing crop testing non-target preservation and a multi-image example testing shape and texture binding.

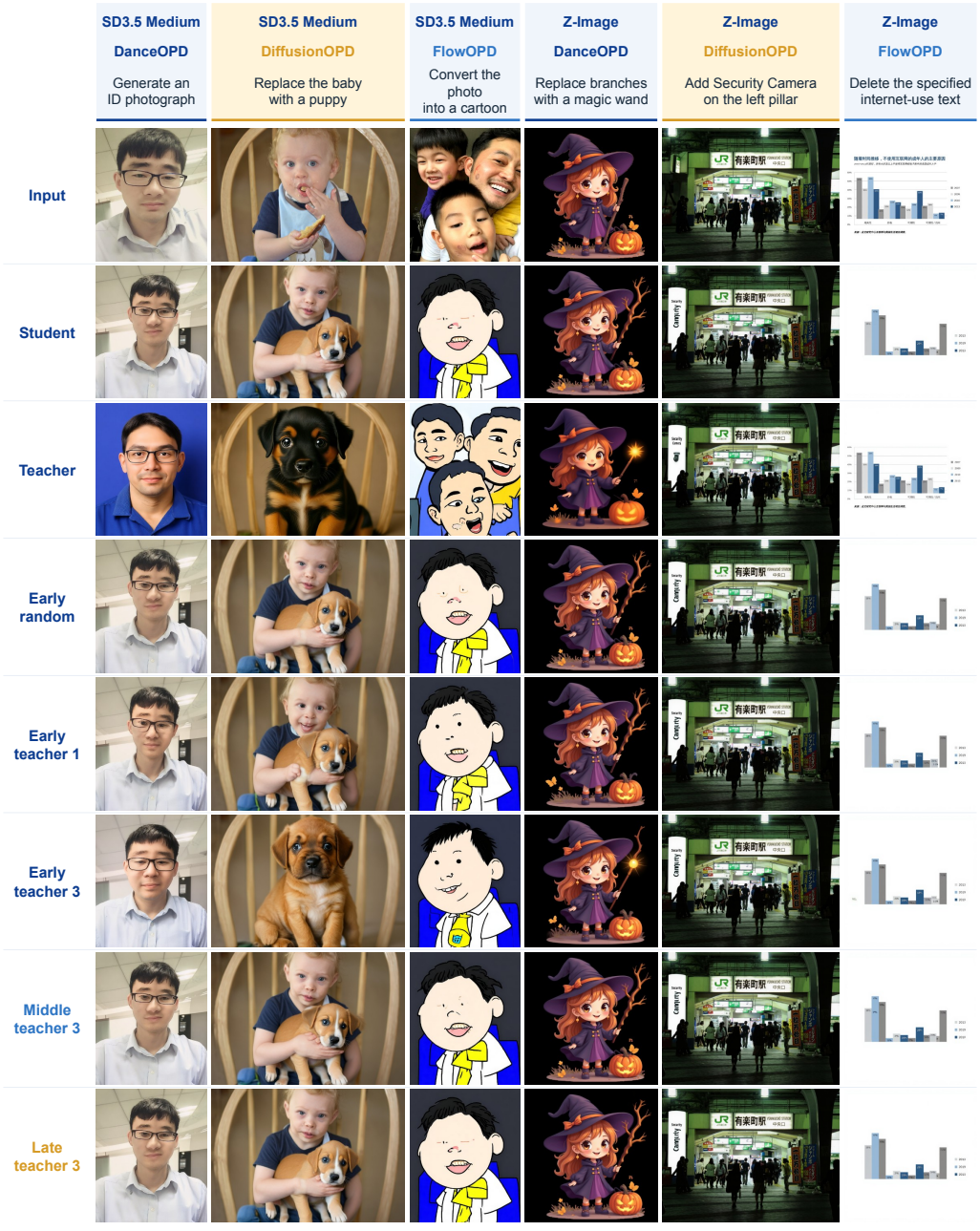

Figure 11. **Qualitative effects of trajectory interventions.** Columns show six matched editing cases; rows compare the input, unmodified student, full teacher, random control, and one- or three-step teacher interventions. All branches continue with the original student.

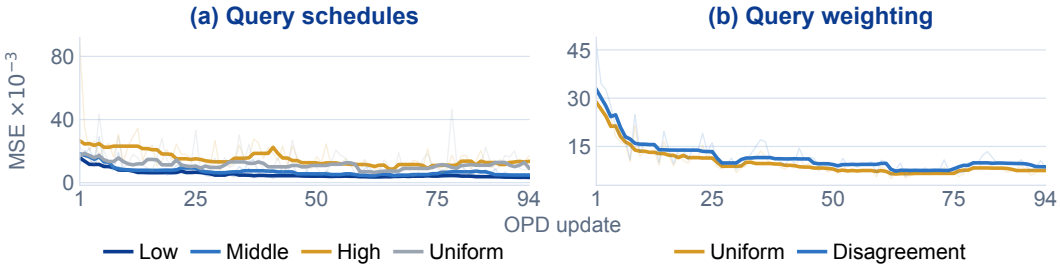

Figure 12. **Training dynamics under temporal supervision.** (a) Velocity-MSE traces for fixed low-, middle-, high-noise, and uniform query schedules. (b) Traces for query-matched uniform and disagreement weighting; faint and bold lines show updates and nine-update medians.

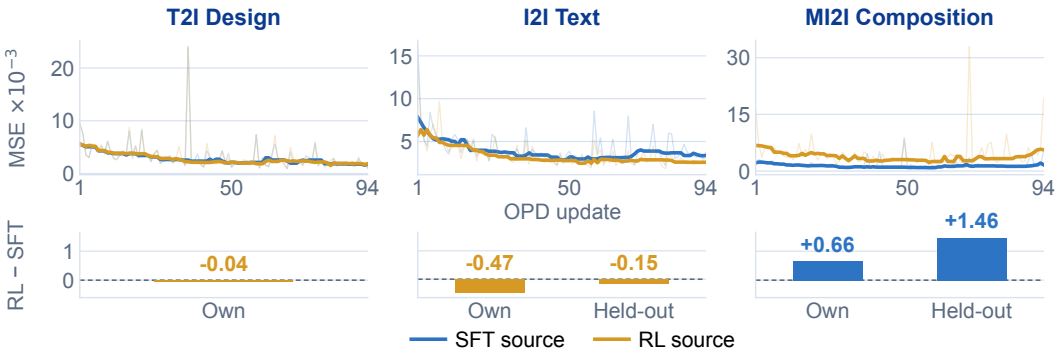

Figure 13. **Teacher fitting and downstream transfer.** Top: velocity-MSE traces for SFT- and RL-source teachers on T2I design, I2I text, and MI2I composition. Bottom: RL-source minus SFT-source student scores on own and held-out routes.

Figure 14 adds eight matched SFT- and RL-source examples across text editing and multi-image composition. Table 8, not these selected images, determines the aggregate ordering.

F.5 Reward transfer and cross-interface outputs

Figure 15 separates native rewards from OPD losses. Figure 16 shows specialist rewards and their transfer to single-teacher students; these are controls, not joint multi-reward competitors.

Figure 17 shows ten matched prompts for single-reward and joint-reward endpoints. The reward label identifies the specialist; joint OPD uses all three specialists. These examples illustrate, but do not determine, the full test-100 results.

Figures 18 and 19 each show eight conditions, four per reward suite, with matched prompts and, for I2I, sources across endpoints. The two grids use different conditions selected for instruction diversity before inspecting results. T2I-teacher I2I outputs retain the initialization's edit adapter.

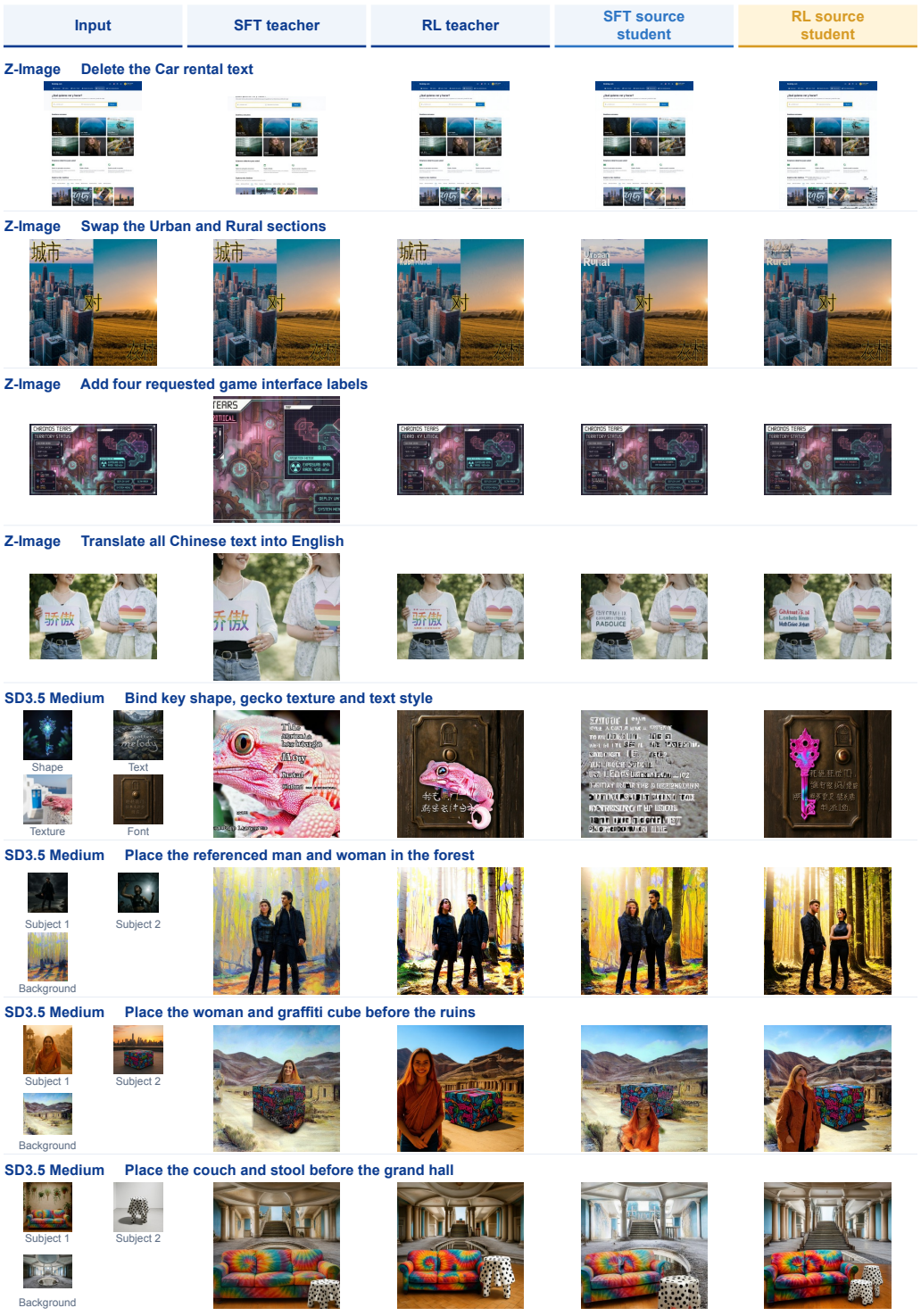

Figure 14. **Qualitative effects of teacher source.** Eight matched text-editing and multi-image conditions compare SFT and RL teachers with their corresponding OPD students, showing differences in requested changes, preservation, binding, and placement.

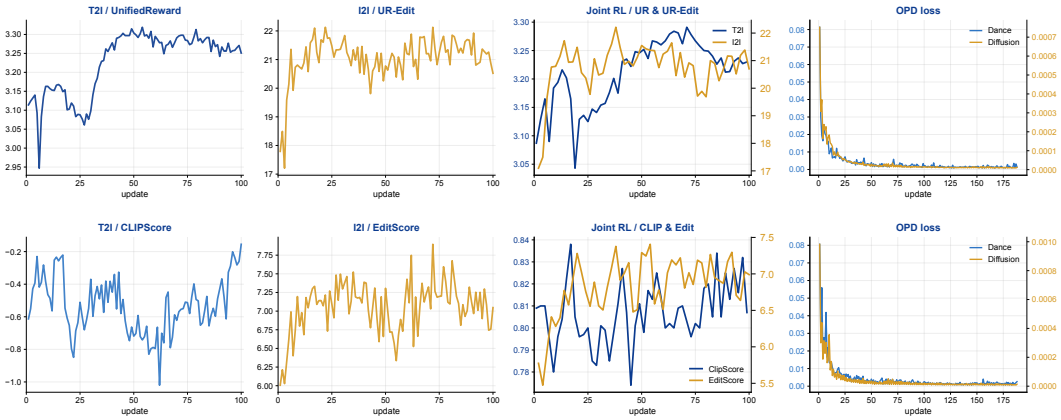

Figure 15. **Training dynamics across two reward suites.** Top: UnifiedReward T2I, I2I, joint-RL, and OPD traces. Bottom: corresponding CLIPScore and EditScore-4B traces. RL panels report native rewards, whereas OPD panels report training loss.

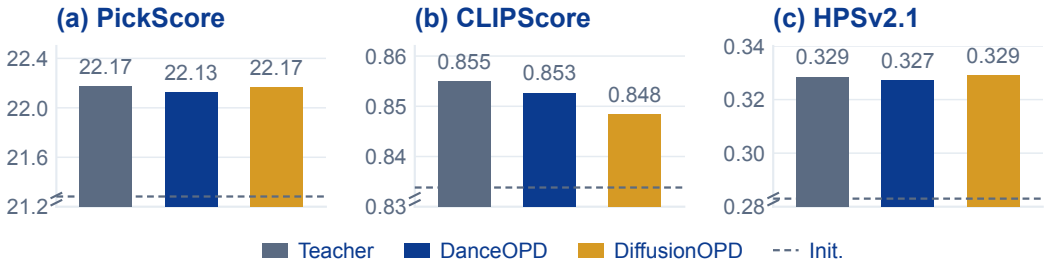

Figure 16. **Single reward teacher-to-student transfer.** PickScore, CLIPScore, and HPSv2.1 panels compare initialization, reward specialists, DanceOPD, and DiffusionOPD.

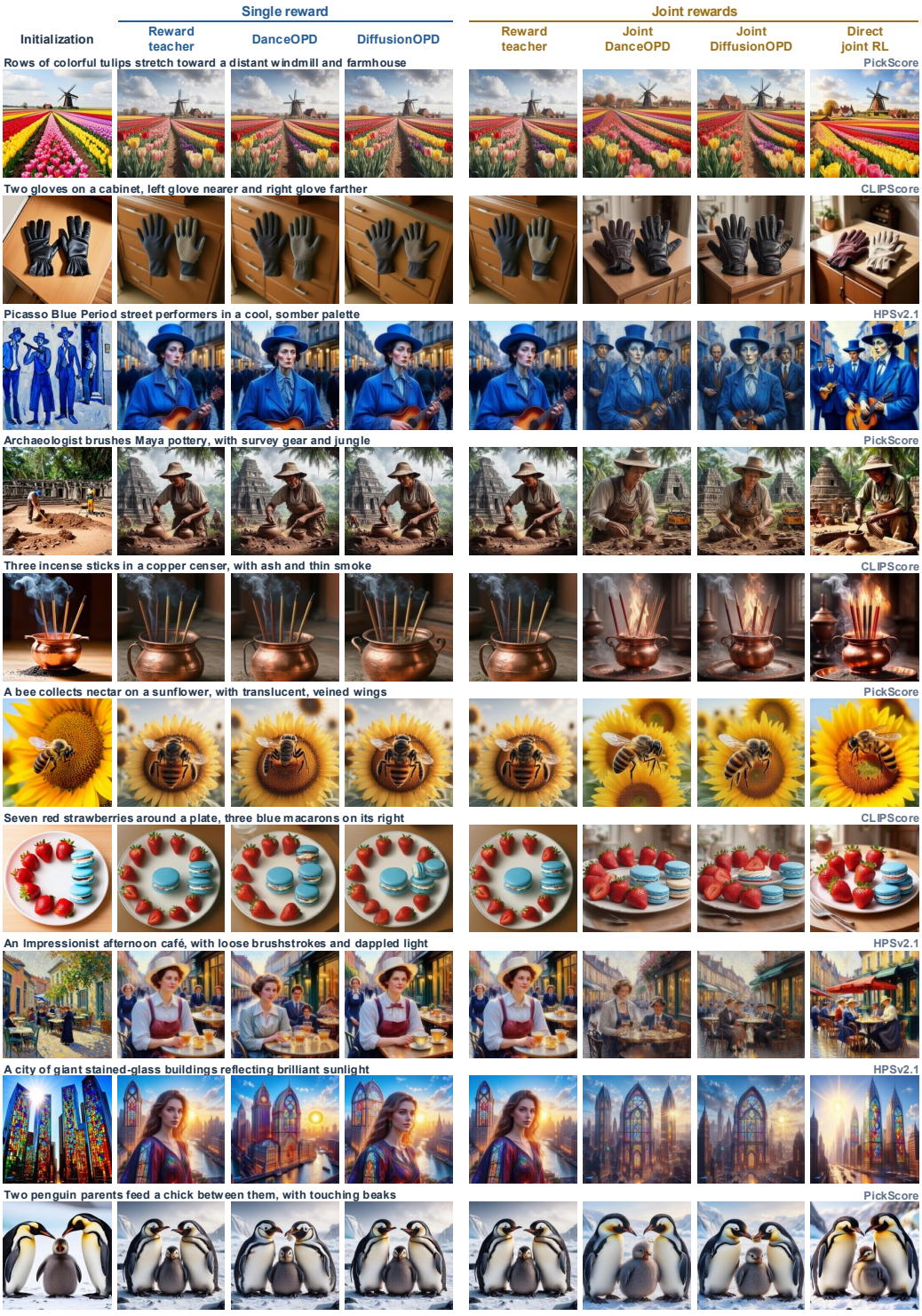

Figure 17. **Single- and joint-reward outputs on matched prompts.**

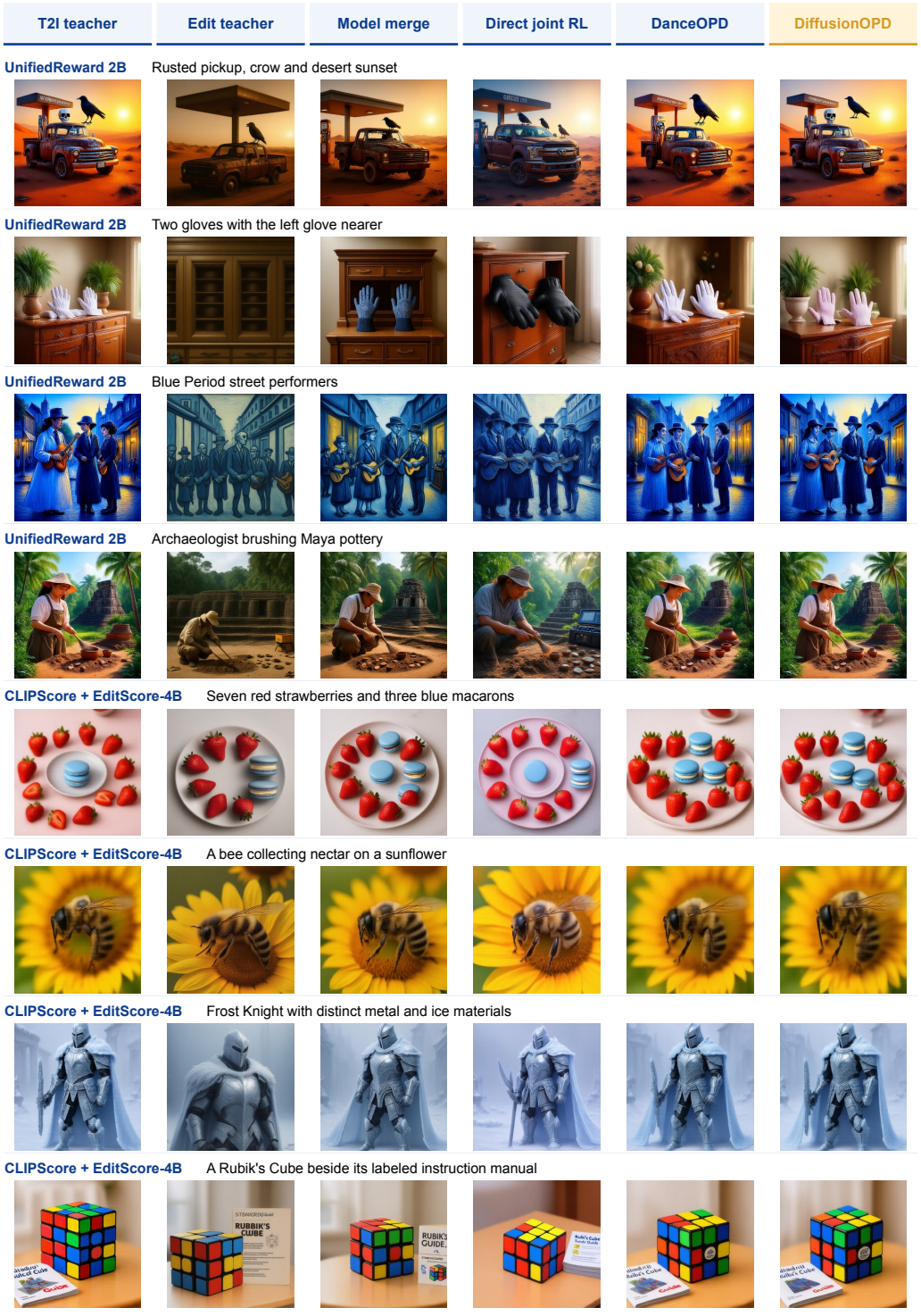

Figure 18. **T2I outputs under two reward suites.** Eight matched prompts compare reward specialists, their parameter midpoint, direct joint RL, DanceOPD, and DiffusionOPD.

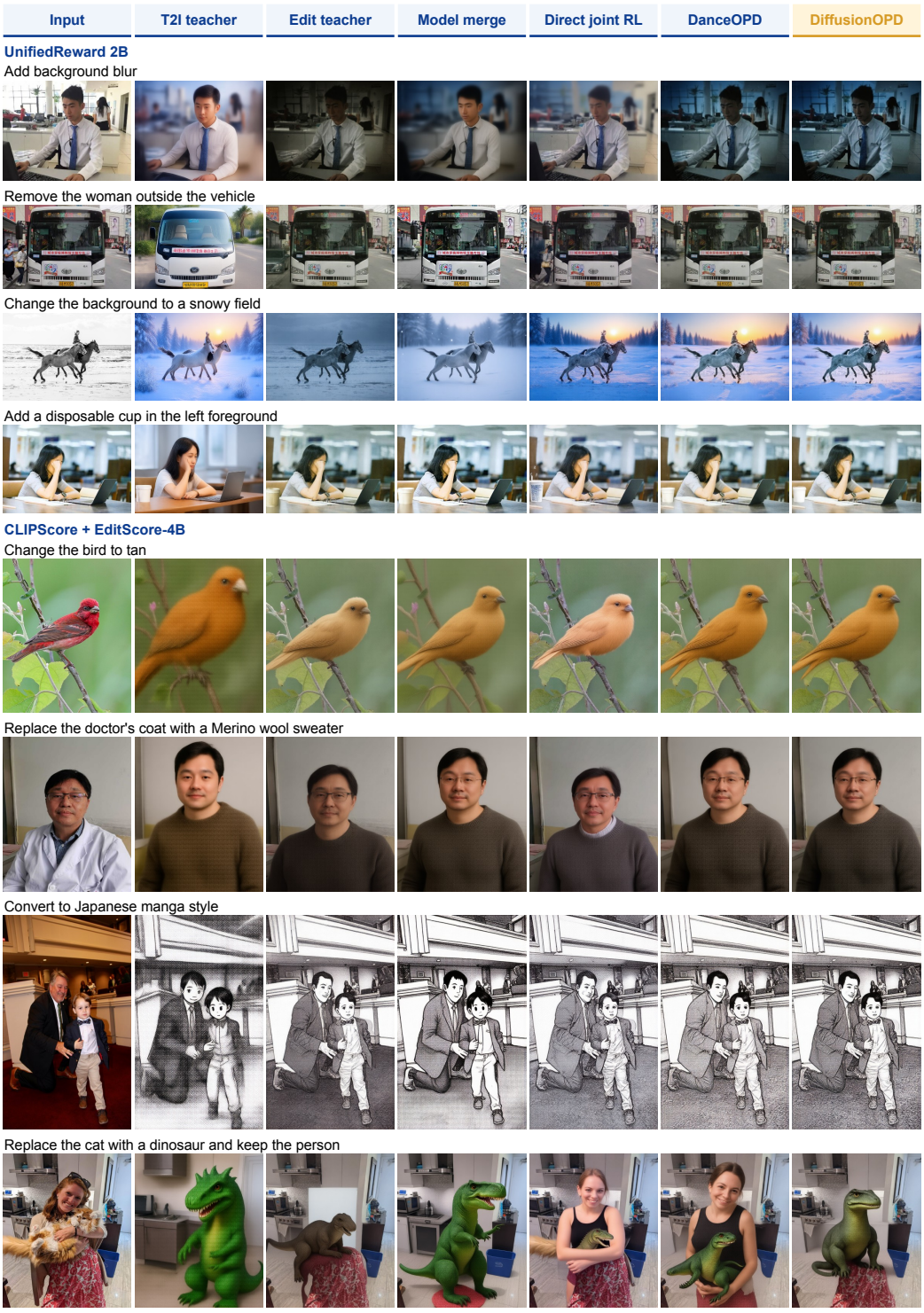

Figure 19. **I2I outputs under two reward suites.** Eight matched source–instruction pairs compare reward specialists, parameter merging, direct joint RL, DanceOPD, and DiffusionOPD.