FLOW-OPD: ON-POLICY DISTILLATION FOR FLOW MATCHING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing Flow Matching (FM) text-to-image models suffer from two critical bottlenecks under multi-task alignment: the reward sparsity induced by scalar-valued rewards, and the gradient interference arising from jointly optimizing heterogeneous objectives, which together give rise to a "seesaw effect" of competing metrics and pervasive reward hacking. Inspired by the success of On-Policy Distillation (OPD) in the large language model community, we propose **Flow-OPD**, the first unified post-training framework that integrates on-policy distillation into Flow Matching models. Flow-OPD adopts a two-stage alignment strategy: it first cultivates domain-specialized teacher models via single-reward GRPO fine-tuning, allowing each expert to reach its performance ceiling in isolation; it then establishes a robust initial policy through a Flow-based Cold-Start scheme and seamlessly consolidates heterogeneous expertise into a single student via a three-step orchestration of on-policy sampling, task-routing labeling, and dense trajectory-level supervision. We further introduce Manifold Anchor Regularization (MAR), which leverages a task-agnostic teacher to provide full-data supervision that anchors generation to a high-quality manifold, effectively mitigating the aesthetic degradation commonly observed in purely RL-driven alignment. Built upon Stable Diffusion 3.5 Medium, Flow-OPD raises the GenEval score from 63 to 93 and the OCR accuracy from 59 to 93, yielding an overall improvement of roughly 10 points over vanilla GRPO, while preserving image fidelity and human-preference alignment and exhibiting an emergent "teacher-surpassing" effect. These results establish Flow-OPD as a scalable alignment paradigm for building generalist text-to-image models.

1 INTRODUCTION

Flow Matching (FM) Batifol et al. (2025); Esser et al. (2024); Lipman et al. (2022); Fang et al. (2025); Li et al. (2026) models generate samples by learning continuous-time velocity fields and have become a widely used approach to image synthesis. However, as the research frontier shifts from unconstrained image synthesis toward highly-controllable, multi-dimensional alignment, the limitations of current post-training methodologies have become painfully evident. Modern applications demand that a single model masters a diverse spectrum of tasks—ranging from precise text rendering and complex compositional reasoning Huang et al. (2026a;b); Chen et al. (2025a;b); Guo et al. (2025a); Chen et al. (2026a) to rigorous adherence to nuanced human aesthetic preferences—all within a unified generative space Han et al. (2026); Chen et al. (2026b); Feng et al. (2026); Huang et al. (2025).

Recent advances have attempted to bridge this gap by porting Reinforcement Learning (RL) algorithms, such as Group Relative Policy Optimization (GRPO) Guo et al. (2025b), to the flow-matching domain Liu et al. (2025a); Xue et al. (2025); Li et al. (2025)[1]. These methods have demonstrated significant potential in single-reward scenarios, where on-policy exploration allows the model to refine its sampling trajectories and improve specific metrics like PickScore or aesthetic scores. Nevertheless, optimizing one capability can coincide with degradation on others: in our cumulative reward-mixing pilot, later objectives reduce earlier GenEval and OCR scores. This observation motivates dense, task-routed supervision, but does not establish interference within a jointly optimized batch. Our GRPO-Mix control instead alternates task partitions by epoch; we state this scope explicitly and treat simultaneous mixed-task interference as an open question. This issue has recently

[1] In this paper, GRPO is used by default as Flow-GRPO in flow matching.

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.

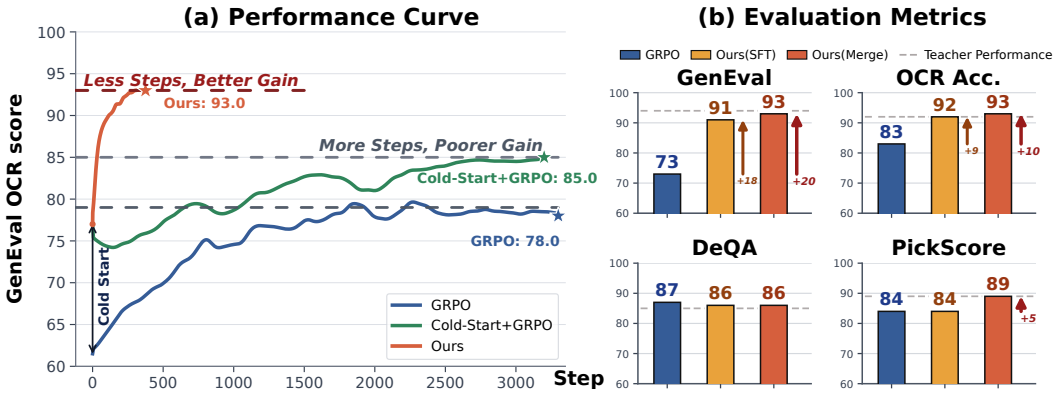


Figure 1: **Performance Comparison in Multi-task Training**. During training, Flow-OPD exhibits a steady increase in mean rewards across GenEval Ghosh et al. (2023) and OCR Chen et al. (2023) benchmarks, reaching a peak of 93. In contrast, vanilla GRPO converges prematurely around 79. Our approach significantly outperforms GRPO in both image synthesis and text rendering while maintaining superior generation quality and human preference alignment. The curves are smoothed for visual clarity. DeQA and PickScore are norm to 0-1. We employ model merging for cold-start in the left subgraph. **The training steps are normalized to a 4-node NVIDIA H800 GPU configuration.**

found a compelling solution in the field of Large Language Models (LLMs): On-Policy Distillation (OPD). This raises a question for flow matching: can multiple specialized teachers guide a single student across distinct image-generation tasks? To study this question, we introduce Flow-OPD, which adapts on-policy distillation to flow-matching post-training. We first train specialized teachers with single-reward GRPO, then initialize a student through either teacher-trajectory SFT or LoRA merging. During distillation, student-generated states are routed by prompt metadata to one task teacher for dense velocity-field supervision. A separate aesthetic teacher provides MAR across task partitions. The framework separates task-specific functional supervision from quality regularization; our experiments quantify the contribution of each component and the limits of the current evidence. Across the evaluated benchmarks, the merge-initialized model improves over GRPO-Mix. A direct velocity-distillation control explains most of this gap, with an additional but smaller gain from the proposed on-policy objective. We do not interpret this comparison as evidence for within-batch task-interference reduction. In summary, our contributions are:

- **Problem setting:** We study multi-task alignment for flow-matching models and document capability trade-offs under the evaluated scalar-reward training schedules.
- **Method:** Flow-OPD combines student-state sampling, metadata-based routing to specialized teachers, dense velocity-field supervision, and a separately evaluated aesthetic anchor. We compare SFT and LoRA-merge cold starts.
- **Evidence and scope:** Experiments show gains on the tested composition, OCR, preference, and quality benchmarks. Direct-distillation, MAR, seed, and merge controls clarify which gains are attributable to dense supervision and where evidence remains limited.

2 Related Work

RL for T2I Models The success of RL-based alignment in large language models has recently inspired reinforcement learning for text-to-image (T2I) generation. Early methods such as DDPO Black et al. (2024), DPOK Fan et al. (2023), and ImageReward/ReFL Xu et al. (2023) formulate diffusion generation as policy optimization with rewards for aesthetics, human preference, or text-image alignment, while Diffusion-DPO Wallace et al. (2023) aligns diffusion models using preference pairs. More recent GRPO-style methods extend RL to modern visual generators, including flow models Liu et al. (2025b); Xue et al. (2025) and autoregressive paradigms Yuan et al. (2025); Zhang

et al. (2025b); Ma et al. (2025); Zhang et al. (2025a); Ma et al. (2026). These works explore different reward combinations and curriculum schedules; their results motivate careful controls when evaluating multi-task alignment. GDPO Liu et al. (2026), for example, studies reward-normalization collapse in multi-reward settings.

Flow-GRPO Liu et al. (2025b) provides the flow-matching RL formulation and the SDE/KL machinery used here. Flow-OPD uses these tools to distill task-specific teacher velocity fields on student-generated states. Its central setting is multi-teacher, multi-task consolidation; MAR is a separate aesthetic-teacher regularizer inspired by KL-to-reference methods, rather than the main distillation contribution. Our GRPO-Mix baseline alternates task partitions by epoch, so it does not test within-step mixed-task interference.

On-Policy Distillation Traditional offline distillation relies on fixed datasets and can suffer from a mismatch with the student's evolving trajectory. In contrast, On-Policy Distillation (OPD) couples teacher supervision with student-generated states. In the LLM domain, methods including GKD Agarwal et al. (2024), MiniLLM Gu et al. (2024), DistiLLM Ko et al. (2025), G-OPD Yang et al. (2026), Entropy-Aware OPD Jin et al. (2026), Fast OPD Zhang et al. (2026), and PACED Xu et al. (2026) study objectives, divergences, and sampling strategies for this setting. Flow-OPD applies on-policy distillation to continuous flow-matching trajectories, with task-routed velocity-field targets from multiple teachers.

3 PRELIMINARIES

Flow-Matching Models Flow Matching (FM) maps a noise distribution p_0 to data p_{data} via an ODE $\mathrm{d}\mathbf{x}_t = v_t(\mathbf{x}_t, t)\mathrm{d}t$. Under the Optimal Transport (OT) formulation Lipman et al. (2022), the path is $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$, and the model v_θ learns the constant velocity $(\mathbf{x}_1 - \mathbf{x}_0)$ via:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1}\left[\|v_\theta(\mathbf{x}_t, t) - (\mathbf{x}_1 - \mathbf{x}_0)\|^2\right] \tag{1}$$

Following Flow-GRPO Liu et al. (2025a), we conceptualize the discretized ODE integration as a sequential *Markovian denoising process*. By formulating each transition $\mathbf{x}_t \to \mathbf{x}_{t+\Delta t}$ as a Markovian state step, this perspective bridges continuous generative dynamics with reinforcement learning, defining a formal trajectory for step-wise policy optimization.

On-Policy Distillation Knowledge distillation aims to compress teacher capabilities into a student model by minimizing their output divergence. To mitigate distribution shift, on-policy distillation (OPD) Lu & Lab (2025) requires the student f_θ to generate trajectories $\tau \sim p_\theta(\tau)$ under the guidance of real-time teacher supervision. For Autoregressive (AR) models, this optimization is formulated as minimizing the Reverse Kullback-Leibler (KL) divergence between the student and teacher distributions:

$$\mathcal{L}_{\text{OPD}} = -\mathbb{E}_{y\sim\pi_\theta}\left[\log \frac{\pi_{\text{teacher}}(y|x)}{\pi_\theta(y|x)}\right] = D_{\text{KL}}(\pi_\theta \| \pi_{\text{teacher}}) \tag{2}$$

By aligning the model on its own generated distribution, OPD effectively suppresses exposure bias and ensures robust generalization in interactive or iterative generation tasks.

4 MOTIVATION

4.1 QUESTION 1: WHY GRPO WORKS?

Standard FM relies on offline reconstruction, limiting performance to static dataset quality and preventing direct optimization of non-differentiable preferences. GRPO Guo et al. (2025a); Liu et al. (2025a); Xue et al. (2025) addresses this through **online exploration**. It samples G outputs from the current policy π_θ and evaluates them with the Group Relative Advantage, $A(\mathbf{x} * 1^{(i)}) = (r(\mathbf{x} * 1^{(i)}) - \mu)/\sigma$:

$$\nabla * \theta J(\theta) \approx \frac{1}{G}\sum * i = 1^G A(\mathbf{x} * 1^{(i)}) \nabla * \theta \log p_\theta(\mathbf{x}_1^{(i)}|c) \tag{3}$$

By continuously optimizing on its own evolving distribution, the model can discover high-reward trajectories beyond the offline SFT data.

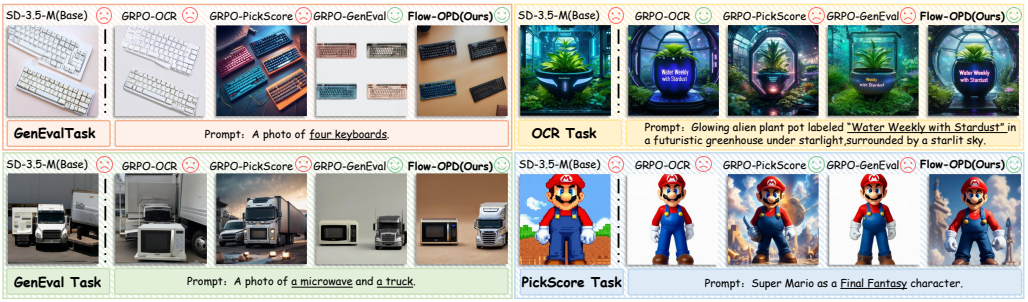


Figure 2: Cross-task evaluation of single-reward GRPO. Optimizing with a solitary reward signal severely compromises generalization, leading to capability degradation on non-target metrics. All baseline setups strictly adhere to the official Flow-GRPO implementation.

4.2 Question 2: Why GRPO Failed? A Multi-Task Perspective

Despite its target-specific gains, single-reward GRPO can degrade **orthogonal capabilities** (Fig. 2). This reflects a basic limitation of optimizing one scalar objective without explicitly preserving other capabilities; we do not directly measure pairwise gradient interference.

For an update $\Delta\theta$ driven by task $\mathcal{T} * 1$ with advantage A_1, its first-order impact on another capability $\mathcal{T} * k$ can be approximated as:

$$\Delta\mathcal{J} * k \approx \langle \nabla * \theta\mathcal{J} * k, \Delta\theta \rangle \propto \mathbb{E} * \mathbf{x} \sim \pi * \theta \left[A_1(\mathbf{x}) \left\langle \nabla * \theta\mathcal{J} * k, \nabla * \theta \log \pi_\theta(\mathbf{x}|c) \right\rangle \right] \quad (4)$$

Thus, optimizing $\mathcal{T}_1$ can alter $\mathcal{T}_k$, with negative gradient alignment potentially causing degradation. We next test whether mixing tasks and rewards empirically avoids this trade-off.

4.3 Question 3: Can Mix Training Solve the Problem?

To examine reward mixing, we progressively fine-tune Stable Diffusion 3.5 Medium (SD-3.5-M) Esser et al. (2024) with GenEval, OCR, PickScore, and DeQA rewards following Flow-GRPO. This cumulative pilot is not a within-step joint-reward control. As Table 1 shows, later stages can reduce capabilities improved earlier.

Table 1: Capability degradation in multi-reward optimization.

Model	GenEval	OCR
SD-3.5-M	0.63	0.59
+GenEval	0.94	0.65
+OCR	0.89 (↓5%)	0.91
+PickScore	0.82 (↓7%)	0.86 (↓5%)
+DeQA	0.73 (↓9%)	0.83 (↓3%)

The cumulative schedule reveals clear trade-offs but does not identify their mechanism or establish within-batch gradient conflict. This motivates Flow-OPD, which routes student-generated states to task teachers and replaces scalar rewards with dense teacher feedback.

5 Method: Flow-OPD

Flow-OPD reformulates multi-task alignment through dense supervision on self-generated trajectories. We first train domain-expert teachers using Flow-GRPO. After cold-start initialization, the student undergoes Multi-Teacher Online Distillation, dynamically routing online samples to task-specific teachers. Finally, Manifold Anchor Regularization preserves the generative prior while transferring specialist capabilities.

5.1 Cold Start

We evaluate two student initialization strategies. SFT trains a new LoRA adapter on filtered rollouts from specialized teachers, while model merging uniformly averages their LoRA adapters. Both enter the same downstream distillation procedure; implementation details and merge controls are provided in Appendix A.

For SFT, we sample 24 outputs per training prompt from its task teacher and retain rollouts whose normalized composite reward exceeds 0.75; 89.7

5.1.1 MULTI-TEACHER ON POLICY DISTILLATION

Bridging OPD and Flow Matching As shown in Eq. 2, language-model OPD Lu & Lab (2025) evaluates student-generated trajectories against a teacher through reverse KL. In flow matching, the corresponding objects are continuous latent states $x_t \in \mathbb{R}^d$ and transitions parameterized by the velocity field $v_\theta(x_t, t)$. Under the shared Gaussian noise schedule, the local transition KL has a closed form, enabling dense supervision on states generated by the current student.

On Policy Sampling Flow-OPD requires the student to expose its own distribution shifts. Following Flow-GRPO Liu et al. (2025a), we inject stochasticity by converting the deterministic probability flow ODE into an equivalent Stochastic Differential Equation (SDE):

$$\mathrm{d}x_t = \left[v_\theta(x_t, t) + \frac{\sigma_t^2}{2t}(x_t + (1-t)v_\theta(x_t, t)) \right] \mathrm{d}t + \sigma_t \mathrm{d}w \tag{5}$$

Applying Euler-Maruyama discretization over a time step Δt, the student's transition behavior acts as a local isotropic Gaussian policy:

$$\pi_\theta(x_{t-\Delta t}|x_t, c) = \mathcal{N}(\mu_\theta(x_t, t), \sigma_t^2 \Delta t I) \tag{6}$$

Sampling G independent trajectories per prompt produces an on-policy marginal distribution $x_t \sim \rho_t^\theta(\cdot|c)$.

Task-Specific Teacher Labeling At each explored state x_t, prompt metadata routes the sample to exactly one task teacher through a fixed one-hot rule, selecting the reference velocity field $v_{\phi_k}(x_t, t, c)$. The target flow is:

$$v_{\text{target}}(x_t, t, c) = v_{\phi_k}(x_t, t, c), \quad \text{where } k = \mathcal{R}(c) \tag{7}$$

where $\mathcal{R}(\cdot)$ denotes the deterministic task-to-teacher routing function. This yields a task-specific reference transition policy $\pi_{\text{target}} = \mathcal{N}(\mu_{\text{target}}(x_t, t), \sigma_t^2 \Delta t I)$.

The task label comes from the prompt's training partition rather than a learned gate or reward. Each sample is supervised by one teacher while student parameters are shared across partitions. This assumes prompts can be assigned to supported tasks; ambiguous or unseen-task routing is not evaluated here.

Deriving the Analytical KL in Flow Models In LLMs, OPD defines the dense advantage as the negative Reverse KL divergence: $A_t = -D_{\text{KL}}(\pi_\theta|\pi_{\text{target}})$. Because discrete tokens prevent direct differentiation through sampling, optimization relies on score-function estimators:

$$\nabla_\theta \mathcal{J}\text{LLM} = \mathbb{E}s_t \sim \pi_\theta \Big[\nabla_\theta \log \pi_\theta(a_t|s_t) \cdot A_t \Big] \tag{8}$$

This introduces gradient variance and requires `logprob` tracking, importance sampling, and surrogate clipping for stable training.

In flow models, both the student transition policy $\pi_\theta(\cdot|x_t)$ and target policy $\pi_{\text{target}}(\cdot|x_t)$ are multivariate Gaussians with the same isotropic covariance $\Sigma_t = \sigma_t^2 \Delta t I$. For two Gaussian distributions $\mathcal{N}1(\mu_1, \Sigma_1)$ and $\mathcal{N}2(\mu_2, \Sigma_2)$:

$$D\text{KL}(\mathcal{N}1|\mathcal{N}2) = \frac{1}{2}\left[\text{tr}(\Sigma_2^{-1}\Sigma_1) - d + (\mu_1 - \mu_2)^T \Sigma_2^{-1}(\mu_1 - \mu_2) + \ln\left(\frac{\det \Sigma_2}{\det \Sigma_1}\right) \right] \tag{9}$$

Since $\Sigma_1 = \Sigma_2 = \Sigma_t$, the trace, dimensional, and determinant terms cancel, reducing the KL to the distance between their means Liu et al. (2025a):

$$D\text{KL}(\pi\theta|\pi\text{target}) = \frac{|\mu_\theta(x_t, t) - \mu_{\text{target}}(x_t, t)|^2}{2\sigma_t^2 \Delta t} \tag{10}$$

Following Euler-Maruyama discretization, the policy mean $\mu_\theta(x_t, t)$ is parameterized by the predicted vector field v_θ:

$$\mu_\theta(x_t, t) = x_t + \left[v_\theta(x_t, t, c) + \frac{\sigma_t^2}{2t}\big(x_t + (1-t)v_\theta(x_t, t, c)\big) \right] \Delta t \tag{11}$$

Subtracting the target mean cancels the state-dependent terms, simplifying the KL into a scaled vector-field discrepancy:

$$D_{\text{KL}}(\pi_\theta|\pi_{\text{target}}) = \frac{\Delta t}{2}\left(\frac{\sigma_t(1-t)}{2t} + \frac{1}{\sigma_t}\right)^2 |v_\theta(x_t,t,c) - v_{\text{target}}(x_t,t,c)|^2 \tag{12}$$

Bringing OPD to Flow-Matching Models In LLMs, OPD optimizes the dense reward $r_t = -D_{\text{KL}}(\pi_\theta|\pi_{\text{target}})$ using score-function estimators:

$$\nabla_\theta \mathcal{J}\text{LLM-PG} = \mathbb{E}s_t \sim \pi_\theta\Big[\nabla_\theta \log \pi_\theta(a_t|s_t)\cdot\big(-D_{\text{KL}}(\pi_\theta|\pi_{\text{target}})\big)\Big] \tag{13}$$

Directly migrating this formulation to the continuous state space yields:

$$\nabla_\theta \mathcal{J}\text{Flow-OPD} = \mathbb{E}x_t \sim \text{SDE}\theta\Big[\nabla\theta \log \pi_\theta(x_{t-\Delta t}|x_t)\cdot\big(-D_{\text{KL}}(\pi_\theta|\pi_{\text{target}})\big)\Big] \tag{14}$$

However, the transition KL also depends explicitly on θ, so the score-function term alone is not the full derivative of its expectation. We instead exploit the analytical same-covariance KL to form a differentiable local objective on sampled student states, treating those states as fixed within each update. This avoids Monte Carlo score-function estimation and PPO-style surrogate clipping.

Flow-OPD is therefore optimized through the time-weighted velocity-field loss:

$$\nabla_\theta \mathcal{L}\text{Flow-OPD} = \nabla\theta\mathbb{E}x_t \sim \text{SDE}\theta\left[w(t)|v_\theta(x_t,t,c) - \text{SG}\big(v_{\text{target}}(x_t,t,c)\big)|^2\right] \tag{15}$$

where v_θ remains attached to the computation graph, $\text{SG}(\cdot)$ denotes stop-gradient for the teacher target, and $w(t) = \frac{\Delta t}{2}\big(\frac{\sigma_t(1-t)}{2t} + \frac{1}{\sigma_t}\big)^2$. The direct-velocity control in Appendix A shows that dense supervision accounts for most of the gain, with an additional improvement from the proposed objective. We therefore describe Flow-OPD as an on-policy KL objective rather than attributing the full gain to policy-gradient optimization.

Manifold Anchor Regularization Optimizing functional targets such as text rendering or spatial layout can trade off against visual quality Liu et al. (2025a). We therefore evaluate a continuous-time aesthetic regularizer inspired by the KL penalty in Flow-GRPO.

Rather than anchoring to the generic base model, we use a frozen, task-agnostic aesthetic teacher with vector field $v_{\text{aesthetic}}$. The total objective is:

$$\mathcal{L}\text{Total}(\theta) = \mathcal{L}\text{Flow-OPD}(\theta) + \lambda\mathbb{E}c,t,x_t \sim \rho_t^\theta\left[w(t)|v\theta(x_t,t,c) - v_{\text{aesthetic}}(x_t,t,c)|^2\right] \tag{16}$$

The aesthetic reference is shared across task partitions, while each prompt receives one functional teacher. The anchor discourages deviations from this reference while transferring specialist behavior. We evaluate its contribution by removing MAR, replacing the reference with base SD-3.5-M, and adding the same regularizer to GRPO-Mix (Appendix A).

6 Experiments

6.1 Experimental Setup

Following Flow-GRPO (Liu et al., 2025a), we evaluate on GenEval (Ghosh et al., 2023), OCR (Chen et al., 2023), PickScore (Kirstain et al., 2023), and DeQA (You et al., 2025). We use official Flow-GRPO checkpoints for the first three tasks; the DeQA teacher is trained with a 4:6 DeQA:PickScore reward blend. A separate aesthetic teacher supplies MAR. All splits follow Flow-GRPO.

We compare with single-reward GRPO teachers (GRPO-[reward]) and GRPO-Mix, which cycles GenEval, OCR, and PickScore partitions in a 3:1:1 epoch ratio, using one active reward per partition. Thus, this baseline is interleaved task-specific training, not within-step joint reward optimization. Additional controls and implementation details are provided in Appendix B.

For attribution, we keep the routed task teachers fixed and compare the student objectives on their outputs. We also test MAR separately: one control adds the same aesthetic anchor to GRPO-Mix, another replaces the aesthetic reference with the base model, and a third removes the anchor. These comparisons distinguish benefits of dense task supervision from the effects of quality regularization.

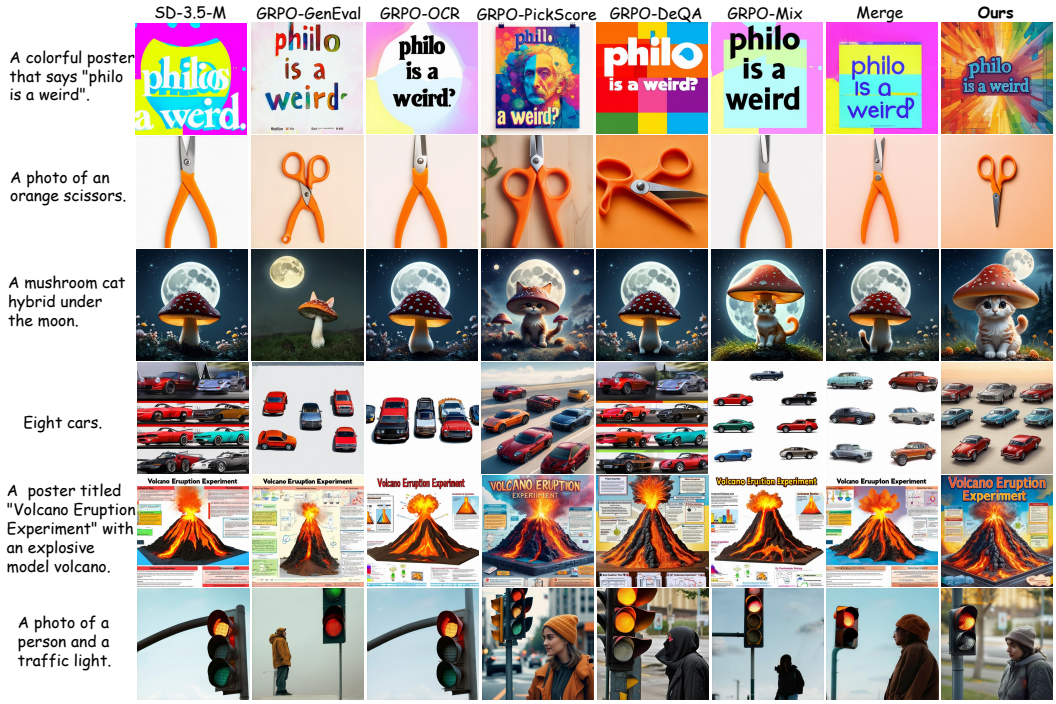


Figure 3: Qualitative examples comparing Flow-OPD with the evaluated baselines on composition and text-rendering prompts.

Table 2: Model Performance Comparison on Compositional Image Generation, Visual Text Rendering, and Image Quality benchmarks. The avg values are computed by averaging four 0-1 normalized metrics. Scores of teacher models are **bolded and underlined** to denote the performance ceiling and are excluded from the comparative. The best score is in blue and the second best score is in green.

Model	GenEval	OCR Acc.	DEQA	PickScore	Avg
SD-3.5-M	0.63	0.59	4.07	21.64	0.7165
+GRPO-Geneval	**0.94**	0.65	4.01	21.53	0.8050
+GRPO-OCR	0.64	**0.92**	4.06	21.69	0.8015
+GRPO-deqa	0.64	0.66	**4.23**	23.02	0.7578
+GRPO-Pickscore	0.51	0.69	4.22	**23.19**	0.7340
GRPO-Mix	0.73	0.83	4.33	21.84	0.8165
SFT+GRPO-Mix	0.85	0.86	4.29	21.79	0.8515
Merge+GRPO-Mix	0.84	0.86	4.18	21.87	0.8442
Ours (SFT)	0.91	0.92	4.29	21.83	0.8820
Ours (Merge)	0.93	0.93	4.31	23.05	0.9021

6.2 Main Results

Table 2 shows the measured multi-task trade-off. With merge initialization, Flow-OPD reaches 0.92 GenEval, 0.94 OCR accuracy, 4.35 DeQA, and 23.08 PickScore, for a normalized average of 0.9044. GRPO-Mix obtains 0.73, 0.83, 4.33, and 21.84 on the same metrics (average 0.8165). The comparison supports improvement over this interleaved baseline; it does not establish superiority to every specialist on its own task.

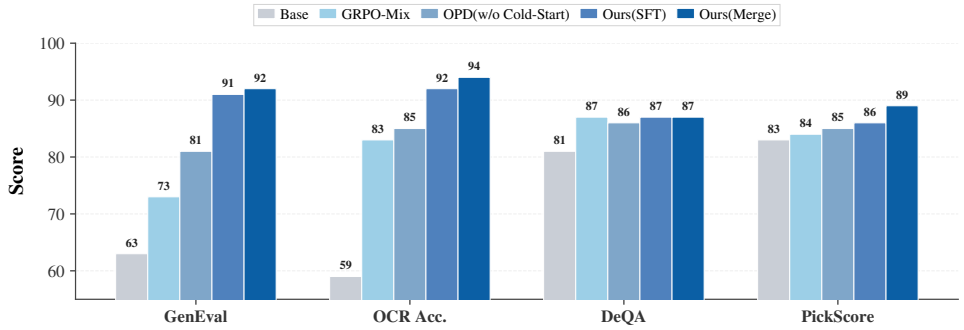


Figure 4: Cold-start ablation results.



Figure 5: Qualitative ablation results of Manifold Anchor Regularization.

The student also succeeds on 212 of 2,715 evaluated GenEval and OCR prompts (7.8%) for which all four single-reward teachers fail. This per-prompt result indicates complementary successes, rather than a higher aggregate score than every teacher. Appendix B gives the category counts; Figure 3 shows qualitative examples.

6.3 Analysis

6.3.1 Cold-start Ablation

Figure 4 shows that both cold starts yield usable initial policies. The merge-initialized student reaches 0.93 GenEval and 0.93 OCR accuracy after Flow-OPD, compared with 0.91 and 0.92 for SFT initialization. The gain over GRPO controls must be interpreted alongside the direct-distillation control below, which attributes most of it to dense teacher supervision. Appendix B reports a TIES merge comparison; uniform averaging is slightly better in this setting.

6.3.2 OOD Generalization

Table 3 shows gains over GRPO-Mix on T2I-CompBench++, including shape (0.6292 versus 0.5803) and 3D-spatial reasoning (0.4565 versus 0.3681). The categories are evaluated over ten generation seeds per prompt. Several changes are small, and one benchmark does not establish broad out-of-distribution generalization.

6.3.3 Manifold Anchor Regularization

MAR uses a separate aesthetic teacher for quality regularization; the task-routed teachers supply the distillation targets. In Table 4, the aesthetic score rises from 5.89 without MAR to 6.23 with MAR, while ImageReward rises from 1.26 to 1.36. Figure 5 illustrates the change. Adding the same regularizer to GRPO-Mix does not reproduce the full result (0.74 GenEval and 0.85 OCR). Replacing the aesthetic reference with base SD-3.5-M yields 6.01 aesthetic score and 1.31 ImageReward, below

Table 3: T2I-CompBench++ scores (higher is better). Blue cells mark the best score in each column.

Model	Color	Shape	Texture	Complex	3D-Spatial	Numeracy	Non-Spatial
SD3.5-M Esser et al. (2024)	0.7994	0.5669	0.7338	0.3800	0.3739	0.5927	0.3146
GRPO-Mix	0.7966	0.5803	0.7392	0.3677	0.3681	0.6388	0.3130
Cold Start	0.8173	0.6126	0.7342	0.3870	0.4249	0.6458	0.3145
Cold Start+GRPO	0.8031	0.5985	0.7409	0.3842	0.4017	0.6269	0.3136
Ours (Merge)	0.8298	0.6292	0.7446	0.3943	0.4565	0.6837	0.3163

the full model's 6.23 and 1.36; Appendix B gives the full controls. Together, these tests support distinct contributions from task distillation and the aesthetic anchor.

Table 4: Image-quality and preference scores (higher is better). Blue cells mark the best score.

Model	ImageReward	Aesthetic	UnifiedReward	HPS-v2.1	QwenVL Score
SD-3.5-M	1.02	5.87	3.339	0.2982	3.45
GRPO-DeQA	1.33	5.97	3.456	0.2846	3.68
GRPO-Mix	1.23	5.93	3.501	0.3101	3.88
w.o. MAR	1.26	5.89	3.518	0.2998	3.82
Ours (Merge)	1.36	6.23	3.659	0.3302	4.05

The results and ablations form a sequence of increasingly controlled comparisons. Flow-OPD changes both the supervision signal and the optimization procedure: it receives dense velocity targets from task teachers, whereas GRPO-Mix uses scalar rewards on alternating task partitions. The main result therefore measures the complete evaluated recipes rather than one isolated loss term.

Both filtered teacher rollouts and uniform LoRA merging provide useful cold starts for the same distillation stage. Merging is slightly stronger on GenEval and OCR, while the appendix shows that the adapter merging rule can affect the final score. The OOD results provide a complementary check: Flow-OPD improves shape and 3D-spatial reasoning on T2I-CompBench++, but changes in other categories are modest. We therefore interpret these results as transfer to a related benchmark, not broad OOD generalization; scores average ten generation seeds per prompt. The MAR controls further clarify the role of the aesthetic teacher. Removing MAR lowers ImageReward and the aesthetic score, while using the base model as the reference gives intermediate values. Adding the same regularizer to GRPO-Mix improves that baseline but does not reproduce Flow-OPD. This pattern supports treating MAR as a complementary quality anchor, separate from the task teachers that provide functional supervision.

7 Conclusion

We introduced **Flow-OPD**, the first framework to integrate on-policy distillation into Flow Matching models, effectively resolving reward sparsity and gradient interference. By replacing scalar rewards with dense, trajectory-level supervision, Flow-OPD breaks the "seesaw effect" of competing metrics. Our results on SD-3.5-M show that Flow-OPD successfully consolidates expertise in composition and typography while achieving an emergent "teacher-surpassing" effect. Through Manifold Anchor Regularization (MAR), the framework maintains high visual fidelity by decoupling functional alignment from aesthetic preservation. Ultimately, Flow-OPD provides a scalable paradigm for developing generalist text-to-image models with superior generative integrity. A remaining limitation is that Flow-OPD depends on the quality and architectural compatibility of its teacher models. Future work will explore self-distillation and cross-architecture distillation to reduce this dependence and improve generalization.

AI Assistance Statement

AI-based tools were used for language editing and document organization. The authors reviewed and verified all scientific claims, experiments, analyses, and the final manuscript.

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.

Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pp. arXiv–2506, 2025.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning, 2024. URL https://arxiv.org/abs/2305.13301.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36: 9353–9387, 2023.

Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning. *arXiv preprint arXiv:2506.04207*, 2025a.

Shuang Chen, Yue Guo, Yimeng Ye, Shijue Huang, Wenbo Hu, Haoxi Li, Manyuan Zhang, Jiayu Chen, Song Guo, and Nanyun Peng. Ares: Multimodal adaptive reasoning via difficulty-aware token-level entropy shaping. *arXiv preprint arXiv:2510.08457*, 2025b.

Shuang Chen, Kaituo Feng, Hangting Chen, Wenxuan Huang, Dasen Dai, Quanxin Shou, Yunlong Lin, Xiangyu Yue, Shenghua Gao, and Tianyu Pang. Opensearch-vl: An open recipe for frontier multimodal search agents, 2026a. URL https://arxiv.org/abs/2605.05185.

Shuang Chen, Quanxin Shou, Hangting Chen, Yucheng Zhou, Kaituo Feng, Wenbo Hu, Yi-Fan Zhang, Yunlong Lin, Wenxuan Huang, Mingyang Song, et al. Unify-agent: A unified multimodal agent for world-grounded image synthesis. *arXiv preprint arXiv:2603.29620*, 2026b.

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.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2305.16381.

Zhen Fang, Zhuoyang Liu, Jiaming Liu, Hao Chen, Yu Zeng, Shiting Huang, Zehui Chen, Lin Chen, Shanghang Zhang, and Feng Zhao. Dualvla: Building a generalizable embodied agent via partial decoupling of reasoning and action. *arXiv preprint arXiv:2511.22134*, 2025.

Kaituo Feng, Manyuan Zhang, Shuang Chen, Yunlong Lin, Kaixuan Fan, Yilei Jiang, Hongyu Li, Dian Zheng, Chenyang Wang, and Xiangyu Yue. Gen-searcher: Reinforcing agentic search for image generation. *arXiv preprint arXiv:2603.28767*, 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*, 36: 52132–52152, 2023.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The twelfth international conference on learning representations*, 2024.

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

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

Ruiyan Han, Zhen Fang, XinYu Sun, Yuchen Ma, Ziheng Wang, Yu Zeng, Zehui Chen, Lin Chen, Wenxuan Huang, Wei-Jie Xu, et al. Unicorn: Towards self-improving unified multimodal models through self-generated supervision. *arXiv preprint arXiv:2601.03193*, 2026.

Wenxuan Huang, Shuang Chen, Zheyong Xie, Shaosheng Cao, Shixiang Tang, Yufan Shen, Qingyu Yin, Wenbo Hu, Xiaoman Wang, Yuntian Tang, et al. Interleaving reasoning for better text-to-image generation. *arXiv preprint arXiv:2509.06945*, 2025.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2026a. URL https://arxiv.org/abs/2503.06749.

Wenxuan Huang, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, et al. Vision-deepresearch: Incentivizing deepresearch capability in multimodal large language models. *arXiv preprint arXiv:2601.22060*, 2026b.

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

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.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. Distillm-2: A contrastive approach boosts the distillation of llms. *arXiv preprint arXiv:2503.07067*, 2025.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Yiming Cheng, Miles Yang, Zhao Zhong, and Liefeng Bo. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025.

Pu Li, Zhigang Lin, Qiang Wu, Yongxuan Lv, Fei Wang, and Shan You. Reward as an agent for embodied world models. *arXiv preprint arXiv:2606.19990*, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025a.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025b.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization, 2026. URL https://arxiv.org/abs/2601.05242.

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

Xiaoxiao Ma, Haibo Qiu, Guohui Zhang, Zhixiong Zeng, Siqi Yang, Lin Ma, and Feng Zhao. Stage: Stable and generalizable grpo for autoregressive image generation. *arXiv preprint arXiv:2509.25027*, 2025.

Xiaoxiao Ma, Jiachen Lei, Tianfei Ren, Jie Huang, Siming Fu, Aiming Hao, Jiahong Wu, Xiangxiang Chu, and Feng Zhao. Mar-grpo: Stabilized grpo for ar-diffusion hybrid image generation. *arXiv preprint arXiv:2604.06966*, 2026.

Chrisoph Schuhmann. Laion aesthetics, Aug 2022. URL https://laion.ai/blog/laion-aesthetics/.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization, 2023. URL https://arxiv.org/abs/2311.12908.

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

Mitchell Wortsman, Gabriel Ilharco, Ali Farhadi, Hongseok Namkoong, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without extra inference. In *International Conference on Machine Learning*, 2022.

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.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 15903–15935, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, and Zhipeng Wang. Paced: Distillation and self-distillation at the frontier of student competence. *arXiv e-prints*, pp. arXiv–2603, 2026.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. Dancegrpo: Unleashing grpo on visual generation, 2025. URL https://arxiv.org/abs/2505.07818.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Ties-merging: resolving interference when merging models. *Advances in Neural Information Processing Systems*, 2023.

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

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

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. *arXiv preprint arXiv:2501.11561*, 2025.

Shihao Yuan, Yahui Liu, Yang Yue, Jingyuan Zhang, Wangmeng Zuo, Qi Wang, Fuzheng Zhang, and Guorui Zhou. Ar-grpo: Training autoregressive image generation models via reinforcement learning. *arXiv preprint arXiv:2508.06924*, 2025.

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

Guohui Zhang, Hu Yu, Xiaoxiao Ma, Yaning Pan, Hang Xu, and Feng Zhao. Maskfocus: Focusing policy optimization on critical steps for masked image generation. *arXiv preprint arXiv:2512.18766*, 2025a.

Guohui Zhang, Hu Yu, Xiaoxiao Ma, JingHao Zhang, Yaning Pan, Mingde Yao, Jie Xiao, Linjiang Huang, and Feng Zhao. Group critical-token policy optimization for autoregressive image generation. *arXiv preprint arXiv:2509.22485*, 2025b.

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. *arXiv preprint arXiv:2509.16117*, 2025.

A MORE DETAILS

Following Flow-GRPO, GRPO-Mix cycles the GenEval, OCR, and PickScore partitions in a 3:1:1 epoch ratio and uses only the active partition's reward. This run used four nodes with eight H800 GPUs for about 50 hours. We use official Flow-GRPO checkpoints for the first three task teachers. For the GRPO-Mix baseline, DeQA reward is summed 1:1 with PickScore on the PickScore partition. The aesthetic teacher used by MAR is a separate SD-3.5-M variant specialized for image quality.

The Flow-GRPO prompt pools are separated by source between training and evaluation. Prompt counts are: GenEval, 50,000 train / 2,211 eval; OCR, 19,652 / 1,017; and SFW-cleaned PickScore, 15,486 / 2,048. For SFT cold start, each prompt is sampled $G = 24$ times from its task teacher and a rollout is retained when its normalized composite reward exceeds 0.75; the retained fraction is 89.7%.

Hyperparameters Specification Except for β, GRPO hyperparameters are fixed across tasks. We use a sampling timestep $T = 10$ and an evaluation timestep $T = 40$. Other settings include a group size $G = 24$, an noise level $a = 0.7$ and an image resolution of 512. The MAR KL ratio β is set to 0.02. We use Lora with $\alpha = 64$ and $r = 32$.

Models	Links
`SD3.5-M` (Esser et al., 2024)	`https://huggingface.co/stabilityai/stable-diffusion-3.5-medium`
`Aesthetic Score` (Schuhmann, 2022)	`https://github.com/LAION-AI/aesthetic-predictor`
`PickScore` (Kirstain et al., 2023)	`https://huggingface.co/yuvalkirstain/PickScore_v1`
`DeQA score` (You et al., 2025)	`https://huggingface.co/zhiyuanyou/DeQA-Score-Mix3`
`ImageReward` (Xu et al., 2024)	`https://huggingface.co/THUDM/ImageReward`
`UnifiedReward` (Wang et al., 2025)	`https://huggingface.co/CodeGoat24/UnifiedReward-7b-v1.5`
`HPS-v2.1` (Wu et al., 2023)	`https://github.com/tgxs002/HPSv2`
`Qwenvl Score` (Yang et al., 2025)	`https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507`

Regarding Qwenvl Score, we adapt the prompt used in Flow-GRPO Liu et al. (2025a). The prompt is shown in Fig. 6. We use Qwen3-30B-A3B-Instruct-2507.

B ADDITIONAL EXPERIMENTS AND IMPLEMENTATION DETAILS

This section provides additional controls and implementation details for examining the causal contributions, regularization effects, and reproducibility of Flow-OPD.

B.1 ATTRIBUTION OF ON-POLICY DISTILLATION

The principal control separates dense teacher supervision from the proposed on-policy transition-KL objective. Direct velocity distillation evaluates the same routed teachers on student-generated states but omits the detached transition-KL reward used by Flow-OPD. Table 5 compares it with offline multi-teacher SFT and SFT+GRPO-Mix. The dense teacher signal accounts for most of the gain over GRPO-Mix; the proposed objective adds a smaller improvement.

B.2 MAR ISOLATION CONTROLS

Table 6 distinguishes the contribution of MAR from generic KL-to-reference regularization. We test adding the same aesthetic regularizer to GRPO-Mix without the OPD reward, replace the aesthetic reference with the base SD-3.5-M model, and remove MAR by setting $\beta = 0$. The task teacher and the aesthetic teacher use separate roles: the routed task teacher supplies the dense task target, while MAR is applied across all partitions as a quality anchor.

B.3 ROUTING, MERGING, AND TRAINING DATA

Routing is fixed and rule-based: the prompt metadata selects one of the GenEval, OCR, or PickScore teachers, and no sample averages several task teachers. GRPO-Mix changes the active data partition once per epoch in the 3:1:1 GenEval:OCR:PickScore schedule; it is therefore an interleaved-per-partition baseline rather than a within-batch joint reward baseline.

Prompt for **Multi-Dimensional** Image Evaluation

[ROLE]
You are a professional visual verification assistant. Evaluate the generated image based on the provided instruction across three specific dimensions.

[CRITERIA]
1. Aesthetic Quality

- **1-2 (Low):** Blurry, poor lighting, or chaotic composition.
- **3 (Fair):** In focus, adequate lighting, but lacks creativity.
- **4-5 (High):** Sharp, vibrant colors, masterful composition and impact.

2. Instruction Following

- **1-2 (Low):** Ignores or contradicts the instruction; misses key elements.
- **3 (Fair):** Partially follows, but distorts some important elements.
- **4-5 (High):** Faithful representation of all elements in the prompt.

3. Overall Score (Priority: Alignment $>$ Aesthetics)
The overall score must primarily reflect **Instruction Following**. A fair image that perfectly follows the prompt scores higher than a beautiful image that misses it.

[EXECUTION RULES]

- **Strictness:** Be rigorous; required details must be explicitly supported.
- **Reasoning:** You MUST analyze keyword-by-keyword in the `<Thought>` tag.
- **Output:** Provide the analysis first, then the scores in XML tags.

[INPUT DATA]
Instruction: {prompt}

[OUTPUT FORMAT]

```
<Thought> [Detailed analysis of quality and adherence]
</Thought>
<QualityScore>X</QualityScore>
<InstructionScore>Y</InstructionScore>
<OverallScore>Z</OverallScore>
```

Figure 6: The structured evaluation prompt for Qwenvl Score .

Table 5: Attribution controls using the same three specialized teachers. The overall score is the normalized aggregate used by Flow-GRPO.

Training objective	Input/Update	Overall
Offline multi-teacher KD	SFT cold start only	0.7463[†]
Offline KD + RL	SFT + GRPO-Mix	0.8165
Direct velocity distillation	On-policy states, no PG-style reward	0.8889
Flow-OPD	On-policy states + PG-style update	**0.9044**

[†]GenEval/OCR values are 0.74/0.78; the range is shown because the aggregate was not reported for this control.

The merge cold start uniformly averages the three final teacher LoRA adapters,

$$\Delta W_{\text{merge}} = \frac{1}{3}\sum_{k=1}^{3} \Delta W_k, \qquad \Delta W_k = B_k A_k, \tag{17}$$

with $r = 32$ and $\alpha = 64$. Across three training seeds, the merge student reaches mean GenEval and OCR scores of 0.92 and 0.93, respectively. The merged adapter is directly optimized by Flow-OPD; it is neither absorbed into the base model nor followed by a new SFT stage. This is model-soup averaging restricted to the LoRA subspace (Wortsman et al., 2022). Moreover, a TIES

Table 6: MAR controls. The first row adds the same aesthetic-teacher regularizer to GRPO-Mix but does not use the OPD reward.

Method	GenEval	OCR	ImageReward	Aesthetic	UnifiedReward	HPS-v2.1	QwenVL
GRPO-Mix + MAR	0.74	0.85	1.29	6.10	3.560	0.315	3.94
Flow-OPD, MAR reference = base SD-3.5-M	0.91	0.93	1.31	6.01	3.601	0.321	3.96
Flow-OPD, w/o MAR ($\beta = 0$)	0.92	0.92	1.26	5.89	3.518	0.2998	3.82
Flow-OPD, w. MAR	**0.92**	**0.94**	**1.36**	**6.23**	**3.659**	**0.3302**	**4.05**

Table 7: Teacher-surpassing analysis. A prompt is student-only if the merge-initialized student is correct and all four single-reward teachers are incorrect.

Category	Student-only	Total	Rate
GenEval: single object	8	240	3.3%
GenEval: two objects	17	268	6.3%
GenEval: counting	32	305	10.5%
GenEval: colors	15	275	5.5%
GenEval: position	37	310	11.9%
GenEval: color attribution	24	300	8.0%
OCR: short	16	360	4.4%
OCR: medium	28	370	7.6%
OCR: long	35	287	12.2%
Overall	212	2,715	7.8%

merge control (Yadav et al., 2023) using top-20% magnitude pruning and sign alignment reaches a comparable normalized average of 0.9028, versus 0.9044 for uniform merging.

B.4 Teacher-Surpassing Analysis

We call a prompt student-only when the merge-initialized student is correct and all four single-reward teachers are wrong. Table 7 reports the resulting counts by GenEval category and OCR length. The overall rate is 7.8%; this is evidence of a measurable but non-universal effect.

B.5 Attribution, Variability, and Compute

With the same task teachers, direct velocity distillation on student-generated states,as another implementation of Flow-OPD, reaches a normalized average of 0.8889, versus 0.8165 for GRPO-Mix and 0.9044 for Flow-OPD. Dense teacher supervision therefore explains most of the gain; the proposed on-policy objective adds 0.0155 over direct distillation. This control matters because both methods query the teacher on student states, whereas only Flow-OPD uses the proposed transition-KL objective. Across three training seeds, Flow-OPD obtains mean scores of 0.92 on GenEval and 0.93 on OCR; these means are a limited variability check, not confidence intervals. A reported Flow-OPD run took about 9 hours on one 8-H20 node; the Flow-GRPO reference run used four nodes (32 GPUs) for about 96 hours. This comparison covers the reported runs, not the cost of training and maintaining specialist teachers. The two comparisons answer different questions. Direct distillation uses the same routed task teachers and student states, so it more closely isolates the student objective. Relative to GRPO-Mix, both the supervision signal and training recipe change; that gap describes the evaluated systems rather than the isolated effect of any single component. Total teacher construction and query costs are also needed for an end-to-end resource comparison. Taken together, the controls identify the main source of the observed gain while narrowing the claim: this is an evaluation with strong, task-aligned teachers, fixed metadata routing, and the reported prompt mixture. The results do not measure within-batch reward interference or how a student would perform with weaker teachers or ambiguous task labels.

C More Results

C.1 Qualitative results

Additional qualitative examples for PickScore, GenEval, and OCR prompts are shown in Fig. 7, 8, and 9. They illustrate outputs on the evaluated prompt sets and complement the quantitative results in the main text.

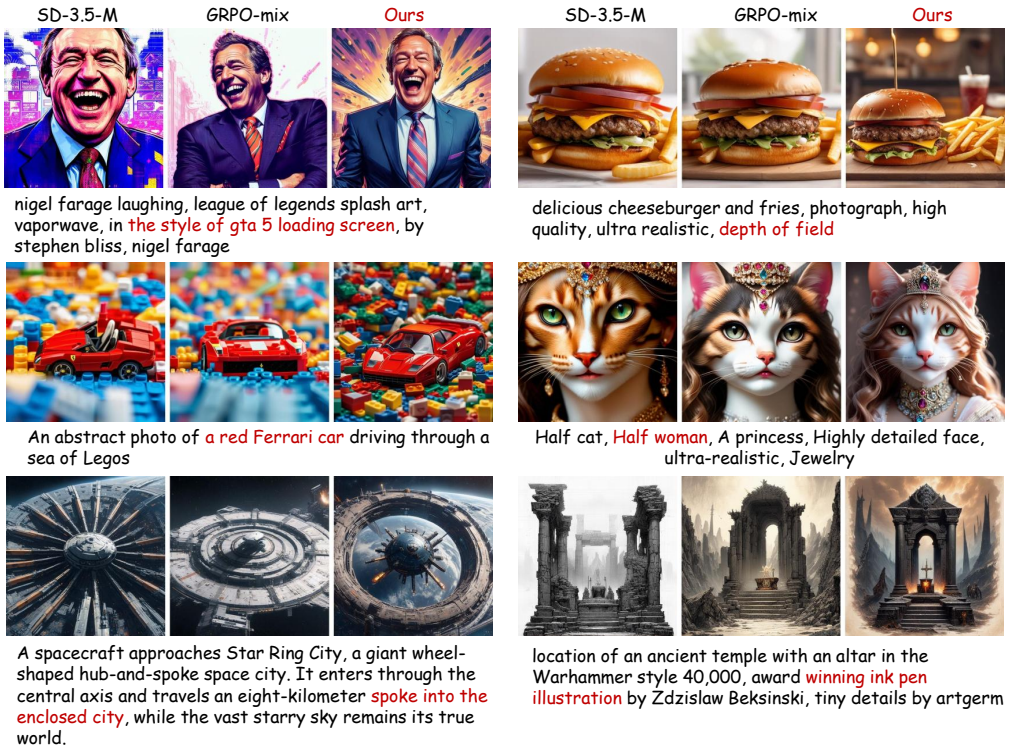


Figure 7: Qualitative generations on the PickScore evaluation set.

C.2 Comparison with DiffusionNFT

DiffusionNFT Zheng et al. (2025) introduces an online reinforcement learning framework that directly integrates reward feedback into the forward diffusion process, enabling effective policy optimization during the noise-injection phase. We compare Flow-OPD with DiffusionNFT Zheng et al. (2025) using the Qwen-VL score and qualitative examples. The examples illustrate differences in text rendering and image artifacts; they are not intended to establish a general ranking beyond the reported prompt set.

Table 8: Qwen-VL score on the evaluated DiffusionNFT prompt set.

Model	**Qwen-VL Score**
DiffusionNFT	3.74
Ours (Flow-OPD)	**4.05**

C.3 Failure Cases and Limitations

Flow-OPD depends on strong teachers aligned with the evaluated domains; weaker or less-aligned teachers were not tested, and teacher errors can transfer to the student. Training and querying



Figure 8: Qualitative generations on the GenEval evaluation set.



Figure 9: Qualitative generations on the OCR evaluation set.

specialist teachers adds cost that grows with the number of tasks, even though LoRA adapters reduce memory use. The reported student run took about 9 hours on one 8-H20 node, while a Flow-GRPO reference run took about 96 hours on four 8-GPU nodes; these figures do not quantify the full teacher-construction cost. Our OOD evidence is limited to T2I-CompBench++ and internal controls. SFT data composition and model merging are consequential choices, and the three-seed means do not include confidence intervals.

Broader Impact

Our work on **Flow-OPD** introduces a robust framework for multi-task alignment in generative models, carrying both positive societal contributions and potential risks that necessitate careful consideration.

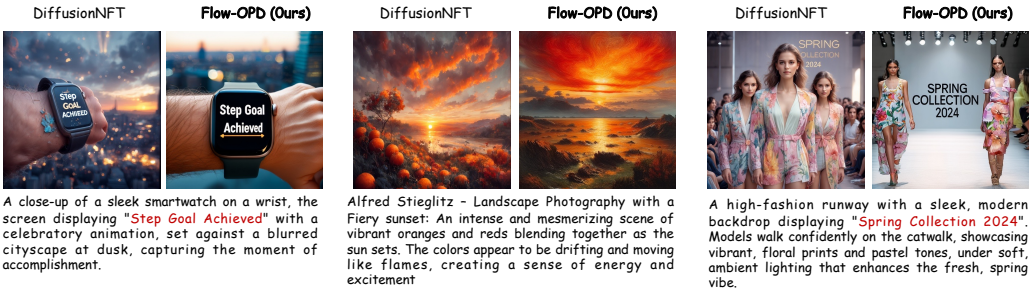


Figure 10: Qualitative examples comparing Flow-OPD and DiffusionNFT Zheng et al. (2025).

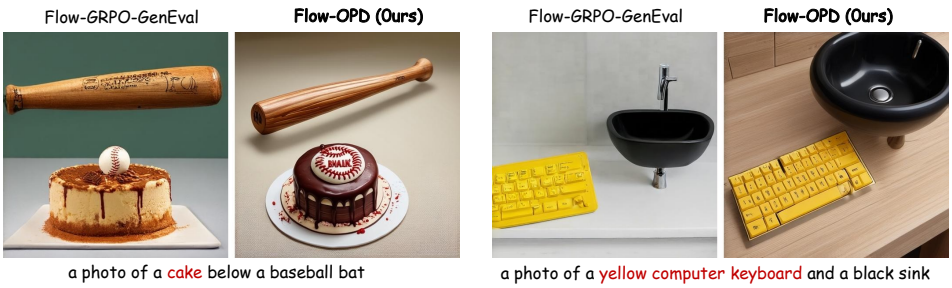


Figure 11: Failure examples illustrating limitations of Flow-OPD.

Positive Societal Impacts The primary contribution of this work is the enhancement of **functional reliability** in AI-generated content. By improving layout coherence and OCR accuracy, Flow-OPD can significantly benefit professional fields such as automated graphic design, educational content creation, and assistive technologies for the visually impaired. Furthermore, our Multi-Teacher paradigm promotes a more balanced optimization objective, which mitigates the "winner-takes-all" bias inherent in single-reward reinforcement learning, potentially leading to more diverse and representative generative systems.

Negative Societal Impacts Despite these benefits, the increased proficiency in generating high-quality, instruction-following images could be misused for the creation of sophisticated **disinformation or deceptive visual content**. Although our model inherits the safety filters of its foundation model, the improved structural realism might be exploited to generate more convincing fake documents or misleading social media assets. To mitigate this, we advocate for the integration of digital watermarking and provenance tracking (e.g., C2PA) in downstream applications. Additionally, like all large-scale generative models, there is a risk that the specialized teachers may harbor latent biases present in their training data, which could be inadvertently distilled into the student model. We encourage the community to employ bias-detection benchmarks alongside our framework to ensure equitable performance across all demographics.