GOJO: GEOMETRY-PRESERVING JOINT OPTIMIZATION FOR ON-POLICY DISTILLATION IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

In post-training of diffusion models, on-policy distillation (OPD) offers an effective way to combine skills from multiple teachers into a single student. While these teachers are obtained through multiple routes, such as supervised fine-tuning (SFT) or reinforcement learning (RL), a unified regression objective is used to match the velocity. Despite its simplicity, this regression objective can lead to unexpected velocity conflicts among teachers' supervision signals and drift in the student model prior, which causes catastrophic forgetting and teacher matching failure. In this paper, we address this limitation through a local probability-reweighting formulation, where we assign credits to each sampled candidate velocity at student-induced states and update the velocities based on the credits. The credits are constructed through three factors: 1) target teacher agreement that directly guides the student velocity; 2) probability redistribution cost that measures how much the student's geometry is perturbed; and 3) mutual teacher damage that quantifies the impact between teachers. We then integrate these three objectives into a single optimization framework, and introduce **GOJO**, *Geometry-preserving Joint Optimization for multi-teacher on-policy distillation*. The resulting optimization problem yields a closed-form solution for the optimal target velocity, which can be used for velocity regression in training. Experiments on SD3.5-Medium and Qwen-Image show that GOJO achieves superior performance in Multi-teacher Distillation with an average 10.5% improvement across tasks on hard-routing OPD and 5.9% on soft-routing OPD, while significantly mitigating the student model prior and avoiding teacher conflicts.

1 INTRODUCTION

Post-training equips diffusion models with specialized capabilities such as image editing, text rendering, and human preference alignment (Wu et al., 2025; Zhao et al., 2026; Xu et al., 2026a; Labs et al., 2025; Seedance et al., 2026). Diffusion post-training commonly consists of three stages: Supervised fine-tuning (SFT) (Ouyang et al., 2022), Reinforcement Learning (RL) (Rafailov et al., 2023; Yue et al., 2025), and On-policy Distillation (OPD) (Agarwal et al., 2024). These approaches reshape the model's conditional generative distribution through different forms of supervision: SFT fits a data-defined target distribution; RL (Liu et al., 2025; Zheng et al., 2025) redistributes probability mass toward higher-reward outputs; and OPD (Li et al., 2026; Zhou et al., 2026) uses teacher supervision on student-induced states to align the student with teacher-defined generative distributions, thereby integrating multiple capabilities into a single model. While the understanding of SFT and RL is rather comprehensive, the OPD methods in diffusion models are less studied. Recent OPD methods, such as DiffusionOPD (Li et al., 2026), DanceOPD (Zhou et al., 2026), and DreOPD (Lin et al., 2026), rely on a practical approach of querying teachers at student rollouts and regressing the student velocity fields to teacher ones.

However, given that a teacher's velocity field contains both desired task-specific capability and other changes introduced by its training data, rewards, and optimization procedure (Zhu et al., 2026a), we argue that ordinary velocity regression has two major limitations: 1) With multiple teachers, regression also transfers changes irrelevant to target capability together, which potentially causes conflicts among teachers. 2) the Euclidean loss does not take care of the student' own geometry, which can lead to prior drift in the student model. We further validate our hypothesis by conducting a two-teacher distillation task on Qwen-Image (Wu et al., 2025; Zhao et al., 2026). We measure the

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: Qualitative comparison of OPD methods. The examples illustrate differences in texture detail and visual style after distilling specialist teachers.

expected velocity and gradient conflict ratios over prompts and denoising steps, and find average 68.9% conflict rates in velocity space and 36.8% in parameter space when teachers are optimized for different objectives. These catastrophic conflicts and uneven teacher matching motivate us to ask:

> *How can we seamlessly transfer teacher capabilities into students while accounting for the student's local probability geometry shift and interference between teachers?*

To answer this question, we view velocity updates through *local probability reweighting* via implicit credits. Specifically, around the student-induced states, we sample multiple candidate velocity update directions and assign each direction a scalar credit. Each credit reflects how a small update in the corresponding direction would affect the desired properties in OPD. Based on these credits, we reweight the distribution over the candidate directions, and the resulting mean shift defines a velocity correction around the student.

Building on this perspective, we introduce **GOJO**, *Geometry-preserving Joint Optimization for multi-teacher on-policy distillation*. GOJO formulates credit assignment as a joint optimization of three criteria: *teacher agreement*, which encourages the proposed velocity to match the relevant teacher; *probability redistribution cost*, which limits the local perturbation to the student's distribution; and *mutual teacher damage*, which penalizes positive increases in other teachers' matching errors relative to the old student. Together, these criteria balance capability transfer with the cost to the student and interference between teachers.

We further show that optimizing over credits can be decomposed into two parts: estimating a covariance matrix representing the student's local geometry and solving for an expected proposal velocity. Thus, for practical training, we first estimate covariance from student-induced forward-velocity residuals using channel, diagonal, or isotropic approximations. Given the estimated covariance, we can then obtain a unique target velocity with a closed-form expression. GOJO then trains the student by regressing directly to this target. The framework accommodates both soft-routing OPD like diffusionOPD (Li et al., 2026) and hard routing methods like DanceOPD (Zhou et al., 2026), where one specialist supplies the task-specific supervision.

Overall, our contributions are threefold: 1) We formulate velocity updates in multi-teacher on-policy distillation as local probability reweighting via implicit credits and show how the student's local covariance determines the minimum quadratic redistribution cost. 2) We introduce GOJO, which jointly optimizes teacher agreement, probability redistribution cost, and mutual teacher damage, with a tractable target solution and practical covariance approximations for both soft- and hard-routing OPD. 3) Experiments on SD3.5-Medium demonstrate average improvements of 10.5% and 5.9% over hard-routing and soft-routing OPD baselines, respectively. Further analysis shows reduced mutual teacher damage and unconditional velocity drift.

2 PRELIMINARIES

Throughout the paper, we work on diffusion-based multi-capability image generation, where we introduce the main components as follows.

Flow Matching (FM). Diffusion models and FM consider fitting a velocity function $v_\theta : \mathcal{S} \to \mathcal{S}$ that iteratively transforms the samples from a source distribution, $x_1 \sim p_1$, to samples in the target data distribution, $x_0 \sim p_0$ (usually accompanied by condition c). Given a diffusion time t, considering the source distribution as isotropic Gaussian $x_1 \sim \mathcal{N}(0, I)$, the samples x_t at t along the trajectory and the corresponding vector field u_t are (Lipman et al., 2023),

$$x_t = \alpha_t x_0 + \sigma_t x_1, \quad u_t = \dot{\alpha}_t x_0 + \dot{\sigma}_t x_1. \tag{1}$$

We consider the rectified flow (Liu et al., 2023) with $\alpha_t = 1 - t$ and $\sigma_t = t$, giving $u_t = x_1 - x_0$. Then, flow matching (Tong et al., 2024) aims at regressing a parameterized velocity $v_\theta(\cdot)$ to u_t by,

$$\mathcal{L}_{\mathrm{CFM}}(\theta) := \mathbb{E}_{(x_0,c)\sim p_0, t, x_1}\left[w(t)\,\|v_\theta(x_t, t, c) - u_t\|^2\right], \quad t \sim U(0, 1) \tag{2}$$

where $w(t) > 0$ is a time-dependent weight. Once the velocity v_θ is properly trained, the generation is conducted through iterative sampling, with $x_{t-dt} = x_t - v_\theta(x_t, t, c)dt$ for $dt > 0$ until $t = 0$.

Post-training of Large-Scale Diffusion Models. Training diffusion model typically involves pre-training to fit the data distribution and post-training to refine the learned distribution for specific objectives. In this paper, we focus on post-training and consider a three-stage recipe comprising *Supervised Fine-Tuning (SFT)*, *Reinforcement Learning (RL)*, and *On-policy Distillation (OPD)*.

SFT adapts a pretrained model and directly shifts the model toward an SFT distribution $p_{\mathrm{SFT}}(x_0, c)$ via an FM-like objective,

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = \mathbb{E}_{(x_0,c)\sim p_{\mathrm{SFT}}, t, x_1}\left[w(t)\|v_\theta(x_t, t, c) - u_t\|^2\right], \tag{3}$$

RL instead trains the model with an advantage function $A(x_0, c)$ using samples collected from a frozen old model, $x_0 \sim p_{\mathrm{old}}$, at the beginning of each update round. With a fixed reference distribution p_{ref}[1], the endpoint objective is (Liu et al., 2025)

$$\max_{p_\theta}\ \mathbb{E}_c\left[\mathbb{E}_{x_0\sim p_\theta(\cdot|c)}[A(x_0, c)] - \beta D_{\mathrm{KL}}\big(p_\theta(\cdot \mid c)\|p_{\mathrm{ref}}(\cdot \mid c)\big)\right]. \tag{4}$$

Its optimum is $p^*(x_0 \mid c) \propto p_{\mathrm{ref}}(x_0 \mid c)\exp\left(A(x_0, c)/\beta\right)$ when the normalizer is finite. One notable RL variant is DiffusionNFT (Zheng et al., 2025), which provides a different route by using a forward-process regression. Writing $h = 0.5 + 0.5\mathrm{clip}(\mathrm{A}/\mathrm{A}_{\max}, -1, 1)$, the loss is

$$\mathcal{L}_{\mathrm{NFT}} = \mathbb{E}\left[h\|v_\theta^+ - u_t\|^2 + (1-h)\|v_\theta^- - u_t\|^2\right], \text{ with } v_\theta^\pm = v_{\mathrm{old}} \pm \gamma(v_\theta - v_{\mathrm{old}}), \gamma > 0 \tag{5}$$

Thus RL supervision can enter through sample-dependent weights h even when the optimized loss is a regression objective. This forward-process property will also be used in our design.

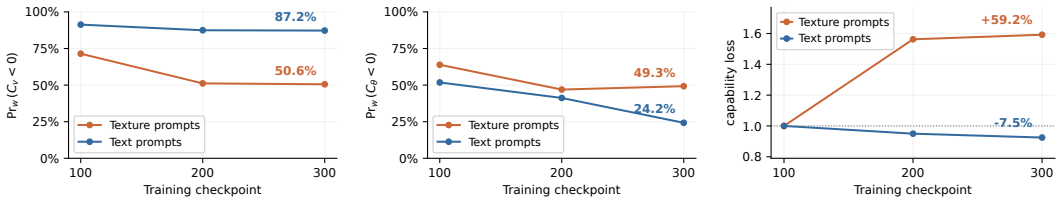
Diffusion On-Policy Distillation (OPD). Recent OPD methods (Li et al., 2026; Zhou et al., 2026) query states induced by the old student ($x_t \sim p_{\mathrm{old}}$), route to a corresponding teacher $k = \mathrm{Route}(c)$, and utilize the teacher-induced field $v_{T_k}(x_t, t, c)$ for distilling the capability, i.e.,

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{c,t,x_t\sim p_{\mathrm{old}}}\left[w(t)\,\|v_\theta(x_t, t, c) - \mathrm{sg}\left[v_{T_k}(x_t, t, c)\right]\|^2\right]. \tag{6}$$

We introduce two notable variants. DiffusionOPD (Li et al., 2026) derives a local mean-matching supervision and applies it densely along student rollout trajectories through t, whereas DanceOPD (Zhou et al., 2026) uses hard routing (i.e. each optimization only queries a single teacher) and trains sparsely only at a single low-noise student-induced state. Despite the differences in query construction, both can be considered as special cases of Eq. (6).

Remark. Eq. (6) depends only on the teacher-supervised velocity field and therefore does not retain how that field was originally obtained (e.g., through SFT, RL, or other sources). In Sec. 3, we show empirically that this potentially causes velocity conflict on the same state and capability degradation within the student model.

[1] p_θ is the model being optimized, p_{old} is its frozen snapshot at the start of the current update round, and p_{ref} is its initial snapshot, kept fixed throughout training.



(a) Velocity conflict rate (b) Gradient conflict rate (c) Teacher matching loss

Figure 2: **Teacher conflict and uneven teacher fidelity during OPD for Qwen-Image.** (a) Mean text-texture velocity cosine C_v averaged over prompts and denoising steps. (b) Gradient cosine C_θ. (c) Source-teacher matching loss normalized to checkpoint 100. Sec. F.2 gives the detailed setting.

3 Why Is Velocity Matching Insufficient?

Depending on the sources, a teacher field used in OPD contains both the desired specialization and other changes introduced by post-training. To understand how the regression objective influences model optimization, we resort to the analysis of post-training in LLMs. Specifically, when consolidating a specific task capability i, Zhu et al. (2026a) decompose the parameter update into two parts: $g_i := \nabla_\theta \log \pi_\theta(x_i) = \mu + \epsilon_i$ with μ the task-irrelevant part and ϵ_i the task-relevant part w.r.t. task i. It is shown that GRPO-like centered advantages (Shao et al., 2024) can remove the former,

$$g_i^{\mathrm{SFT}} = \mu + \bar{\epsilon}_i, \qquad g_j^{\mathrm{RL}} = \frac{1}{G}\sum_{k=1}^{G} \widehat{A}_{j,k}\epsilon_{j,k}, \qquad \sum_{k=1}^{G} \widehat{A}_{j,k} = 0. \tag{7}$$

Here i and j denote different tasks, and $\bar{\epsilon}_i = G^{-1}\sum_k \epsilon_{i,k}$ with G the group size and $\epsilon_{i,k}$ the k-th sample. The RL centralization would cancel the task-irrelevant part μ, while SFT does not. Ideally, the OPD should only distill the capability-relevant part ϵ_i. However, we show that a velocity regression objective would make the shared component enter the RL-teacher update:

Proposition 1. *(Informal) Consider an SFT teacher for task i and an RL teacher for task j, having different prompt routes but distilled into one shared student v_θ with equal weights. One gradient step of size η of OPD and RL, respectively, gives θ_{OPD} and θ_{RL}. With Jacobian $J(x_t^{(i)})$ and Hessian $H_i = \mathbb{E}_{\nu_i}[w(t)J(x_t^{(i)})^\top J(x_t^{(i)})]$, the induced velocities satisfy (derived in Sec. B)*

$$v_{\theta_{\mathrm{OPD}}}(x_t^{(j)}) - v_{\theta_{\mathrm{RL}}}(x_t^{(j)}) = \eta J(x_t^{(j)})H_i\mu + O(\eta^2). \tag{8}$$

where $x_t^{(j)}$ and $x_t^{(i)}$ are queries for task i and j, $v_{\theta_{\mathrm{OPD}}}$ and $v_{\theta_{\mathrm{RL}}}$ are resulting velocity.

Thus in OPD, μ for task i can affect RL task j, even RL can cancel its own shared component.

We also provide an empirical study where we ask a Qwen-Image student jointly learning from texture (RL-induced) and text (SFT-induced) specialists at different training steps. At a fixed query (x_t, t, c), considering the velocity Jacobian as $J_c = \partial v_\theta(x_t, t, c)/\partial\theta$, For $\ell_i = w(t)\|v_\theta - v_{T_i}\|^2$, the gradient descent is $g_i = -\nabla_\theta \ell_i = w(t)J_c^\top(v_{T_i} - v_\theta)$. Using cosine similarity $C_v(i,j) = \cos(v_\theta - v_{T_i}, v_\theta - v_{T_j})$ and $C_\theta(i,j) = \cos(J_c^\top(v_\theta - v_{T_i}), J_c^\top(v_\theta - v_{T_j}))$, we can define velocity conflict C_v and gradient conflict C_θ. Thus, we define the conflict ratio averaged over denoising steps and prompts as the metrics as,

$$R_v = (\sum_{i,j} w_{i,j}\,\mathbf{1}[C_v(i,j) < 0])/(\sum_{i,j} w_{i,j}), \qquad R_\theta = (\sum_{i,j} w_{i,j}\,\mathbf{1}[C_\theta(i,j) < 0])/(\sum_{i,j} w_{i,j}).$$

We evaluate the velocity and gradient conflicts on student rollouts across prompts and denoising steps. The results can be found in Fig. 2. Throughout training, a strong velocity conflict is found, and this conflict leads to a failure of student models to fit the texture teacher. We also visually validate the generated images (see Fig. 1) and find a loss of texture capability in the generation.

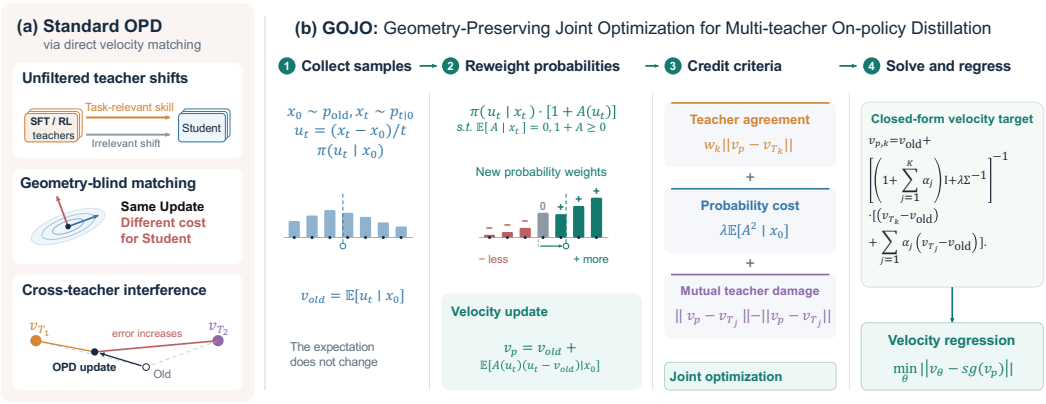


Figure 3: Overview of GOJO. (a) The standard OPD transfers undesired properties into the student model, while velocity regression does not preserve student geometry and causes cross-teacher interference. (b) Pipeline of GOJO. 1. multiple queries are sampled to estimate the local geometry in student-induced states; 2. each velocities candidate are assigned with a credit; 3. The credits define an expected proposal velocity; and 4. the final objective is regressing toward the proposal velocity.

Remark: This section points out two major limitations of direct velocity regression for OPD: 1) With multiple teachers, regression transfers changes irrelevant to the target capability together, which potentially causes conflicts among teachers; and 2) direct velocity regression would cause a prior drift in the student model, which makes some original capability degeneration.

4 GOJO: Geometry-preserving Joint Optimization for Multi-teacher Distillation

From the above analysis, we therefore ask: how should we choose an update that improves teacher agreement for a specific task, while minimizing its negative impact on the student model and other teachers? To this end, we introduce GOJO, which constructs a distillation target by jointly considering teacher agreement, probability redistribution cost, and mutual teacher damage.

4.1 Velocity Updates through Probability Reweighting

A probabilistic view of the student-induced velocity. We first propose a probabilistic approach to interpret the velocity geometry of the student-induced states at an arbitrary time t. Following the forward-process construction in DiffusionNFT (Zheng et al., 2025), we sample x_0 through the student model and x_t via forward renoising $x_t = (1-t)x_0 + tx_1$. Thus, the instance velocity can be written as $u_t = (x_t - x_0)/t$. Given that x_0 is student-induced, sampling multiple x_t enables us to estimate the distribution of $\pi(u_t \mid x_t)$, which represents the student-induced velocity geometry at t. Specifically, we denote the first-order and second-order moments as,

$$\mathbb{E}[u_t \mid x_t] \quad \text{and} \quad \Sigma = \mathbb{E}[(u_t - \mathbb{E}(u_t \mid x_t))(u_t - \mathbb{E}(u_t \mid x_t))^\top \mid x_t],$$

where we assume $v_{\text{old}}(x_t) = \mathbb{E}[u_t \mid x_t]$ and Σ reflects the local geometry around u_t, i.e., how velocities vary around it.

Re-weighting updates via credit. To measure how an instant update direction u_t impacts the overall optimization, we assign each u_t a scalar credit $A(u_t)$. We denote a small u_t-induced update step as $\Delta u = u_t - v_{\text{old}}$, and the reweighted $\pi_A = \pi(u_t \mid x_t)(1+A)$ is still a valid distribution if $\mathbb{E}[A \mid x_t] = 0$. Thus, the mean velocity and probability cost become

$$\mathbb{E}_{\pi_A}[u_t] = v_{\text{old}}(x_t) + \mathbb{E}[A \cdot \Delta u \mid x_t], \qquad D_{\text{KL}}(\pi_A \| \pi) = \mathbb{E}[A^2 \mid x_0] + o(A^3). \tag{9}$$

Appendix C.1.1 provides the derivation. The above two terms measure the local probability cost of changing velocities. The first term defines a proposal velocity field $v_{\text{p}} = v_{\text{old}} + \delta_A$ where

$\delta_A = \mathbb{E}[A(u_t)(u_t - v_{\text{old}}) \mid x_0]$, which is the expected change reweighted by credit configuration A. The second term quantifies the local probability cost of moving from old distribution π to new π_A.

4.2 Distillation as a Joint Optimization Problem

With the above insight, we are two steps away from finding an optimal velocity update. First, we need to quantify what is a good proposal distribution induced by A; second, we need to optimize over the velocity space to obtain the optimal velocity. To this end, we consider the following three factors to quantify the impact of the proposal velocity:

1. **Teacher Agreement.** The proposal velocity update should match the relevant teacher predictions. For teacher k, this is measured by $\|v_{\text{old}} + \delta_A - v_{T_k}\|^2$.
2. **Probability Redistribution Cost.** The update should avoid excessive changes to the student's local distribution. As derived above, $D_{\text{KL}}(\pi_A \| \pi)$ gives the exact cost and $\mathbb{E}[A^2 \mid x_t]$ provides a quadratic approximation of such change in the probabilistic distribution.
3. **Mutual Teacher Damage.** Improving agreement with one teacher may worsen agreement with another. We therefore penalize positive increases in teacher-matching error w.r.t v_{old}. This is denoted as $[\|v_{\text{p}} - v_{T_j}\|^2 - \|v_{\text{old}} - v_{T_j}\|^2]_+, \forall j \neq k$ with $[\cdot]_+ = \max(\cdot, 0)$.

Let w_k and q_k be nonnegative matching and protection weights for each task with $\sum_k w_k = \sum_j q_j = 1$. We can formulate credit assignment problem as

$$\min_A \sum_k w_k \Big[\|v_{\text{old}} + \delta_A - v_{T_k}\|^2 + \lambda \mathbb{E}[A^2 \mid x_0] + \gamma \sum_{j \neq k} q_j \Big[\|v_{\text{old}} + \delta_A - v_{T_j}\|^2 - \|v_{\text{old}} - v_{T_j}\|^2 \Big]_+ \Big] \tag{10}$$

s.t. $\mathbb{E}[A \mid x_t] = 0$ and $\mathbb{E}[A^2 \mid x_t] < \infty$. Here $\lambda, \gamma \geq 0$ control the student redistribution and mutual-teacher protection penalties. The three terms depend on A only through δ_A. Thus, we can define a proposal velocity v_{p} and directly optimize that. Using $v_{\text{p}} := \delta_A + v_{\text{old}}$, we have $\mathbb{E}[A^2 \mid x_t] = (v_{\text{p}} - v_{\text{old}})^\top \Sigma^\dagger (v_{\text{p}} - v_{\text{old}})$ with $\Sigma^\dagger$ the Moore–Penrose pseudoinverse.

Proposition 2. *(Proposal Velocity Optimization Form) We can reformulate Eq.* (10) *and derive the equivalent finite-dimensional problem in optimizing v_{p}(Proof in Sec. C.1.1).*

$$\min_{v_{\text{p}}} \sum_k \Big[\underbrace{\|v_{\text{p}} - v_{T_k}\|^2}_{\text{teacher agreement}} + \underbrace{\lambda (v_{\text{p}} - v_{\text{old}})^\top \Sigma^\dagger (v_{\text{p}} - v_{\text{old}})}_{\text{probability redistribution cost}} + \underbrace{\gamma \sum_{j \neq k} \big[\|v_{\text{p}} - v_{T_j}\|^2 - \|v_{\text{old}} - v_{T_j}\|^2 \big]_+}_{\text{mutual teacher damage}} \Big] \tag{11}$$

For positive definite Σ, $\Sigma^\dagger = \Sigma^{-1}$. We consider uniform matching where $w_k = q_k = 1/K$.

We can obtain a closed-form solution as the objective is strongly convex. (Details in Sec. C.1.3)

Proposition 3. *At the optimal dual coefficients $\{\alpha_k^*\}$, Eq.* (11) *has the unique optimal solution*

$$v_{\text{p}} = v_{\text{old}} + \Sigma \left[\left(1 + \sum_k \alpha_k^* \right) \Sigma + \lambda I \right]^{-1} \cdot \sum_k (w_k + \alpha_k^*)(v_{T_k} - v_{\text{old}}). \tag{12}$$

Once we derive the analytical solution for the optimal velocity, we can rely on a regression for training, i.e., $\mathcal{L}_{\text{GOJO}}(\theta) = \mathbb{E}_{x_t \sim p_{\text{old}}} \Big[w(t) \cdot \|v_\theta(x_t) - \text{sg}\,[v_{\text{p}}(x_t)]\|^2 \Big]$. We further derive three variants: 1) **No Mutual Teacher Damage Term.** Setting $\gamma = 0$ with $\lambda > 0$ gives $\alpha^* = 0$, we obtain $v_{\text{p}} = v_{\text{old}} + \Sigma(\Sigma + \lambda I)^{-1} \sum_k w_k (v_{T_k} - v_{\text{old}})$; 2) **No Re-distribution Cost Term.** With geometry disabled $\lambda = 0$, the proposal becomes a convex teacher mixture $v_{\text{p}} = [\sum_k (w_k + \alpha_k^*) v_{T_k}]/(1 + \sum_k \alpha_k^*)$. 3) **Both eliminated.** For $\lambda = \gamma = 0$, the objective simply recovers the velocity target of soft-routing OPD (Li et al., 2026), i.e. $v_{\text{p}} = \sum_k w_k v_{T_k}$.

Model	Wall-clock time	Rule-Based		Model-Based					
		GenEval	OCR	PickScore	ClipScore	HPSv2.1	Aesthetic	ImgRwd	Model-Based Avg
SD-XL‡	—	0.55	0.14	/	0.287	0.280	5.60	0.76	/
SD3.5-L‡	—	0.71	0.68	/	0.289	0.288	5.50	0.96	/
FLUX.1-Dev‡	—	0.66	0.59	/	0.295	0.274	5.71	0.96	/
SD3.5-M (w/o CFG)	—	0.29	0.12	0.76	0.242	0.206	5.61	-0.52	6.31
+ CFG	—	0.64	0.56	0.83	0.281	0.280	5.40	0.82	7.64
GenEval Teacher	-	0.98	0.34	0.79	0.253	0.253	5.73	0.52	7.56
OCR Teacher	-	0.23	0.97	0.74	0.230	0.190	5.40	-0.86	5.70
Aes Teacher (PickScore)	-	0.61	0.67	0.91	0.279	0.328	6.59	1.31	9.41
DiffusionOPD	1.67x	0.66	0.47	0.85	0.270	0.307	6.32	1.00	8.75
DanceOPD	1.00x	0.82	0.92	0.88	**0.288**	0.328	6.29	**1.39**	9.11
GOJO	1.09x	**0.88**	**0.95**	**0.92**	0.273	0.334	**6.71**	1.31	**9.52**

Table 1: **Multi-teacher OPD Results.** Results with ‡ are taken from (Li et al., 2026). Underline means best among teachers and **bold** is the best among OPD methods.

4.3 Practical Implementations

Estimating the Covariance in Probability Geometry. The ideal covariance Σ of dimension $D \times D$ in Eq. (12) is needed to be estimated during training. For images D is typically $C \times H \times W$ with C the number of channels and $P = H \times W$ the spatial positions. A full parameterization of the covariance is commonly intractable due to the tremendous dimensionality; thus we rely on approximations for dimension reduction. We introduce three ways: channel-wise, diagonal, and isotropic covariance. Channel-wise covariance $\Sigma_{\text{ch}} \in \mathbb{R}^{C\times C}$ maintains a scalar covariance for each channel; diagonal covariance does not consider inter-channel covariance and only maintains a vector $\boldsymbol{c} \in \mathbb{R}^C$; isotropic covariance only has a scalar p. Thus, the full covariance can be recovered via:

$$\widehat{\Sigma} \in \left\{ \widehat{\Sigma}_{\text{ch}} \otimes I_P \text{ (channel)},\ \operatorname{diag}(\boldsymbol{c}) \otimes I_P \text{ (diagonal)},\ pI_{CP} \text{ (isotropic)} \right\}. \tag{13}$$

Their estimation and self-consistency limitations are detailed in Appendix E.3 and E.6.

Hard Routing. Zhou et al. (2026) introduces a hard routing mechanism when each optimization epoch only routes to a specific teacher. Similarly, we derive the proposal distribution in the hard-routing regime, where the details can be found in Appendix E.2,

$$v_{p,k} = v_{\text{old}} + [(1 + \sum_{j=1}^{K} \alpha_j)I + \lambda\Sigma^{-1}]^{-1}[(v_{T_k} - v_{\text{old}}) + \sum_{j=1}^{K} \alpha_j (v_{T_j} - v_{\text{old}})]. \tag{14}$$

Queries construction and teacher field parameterization. We generate x_0 with the frozen old student, and the queries x_t are obtained through re-noising as we discussed in Appendix E.1. All teacher, old-student, and trainable-student evaluations use the same query. Our current recipe uses a frozen base-unconditional anchor, differing from teacher self-CFG.

With all the practical designs, we are able to introduce the algorithm, which we detail in algorithm 1.

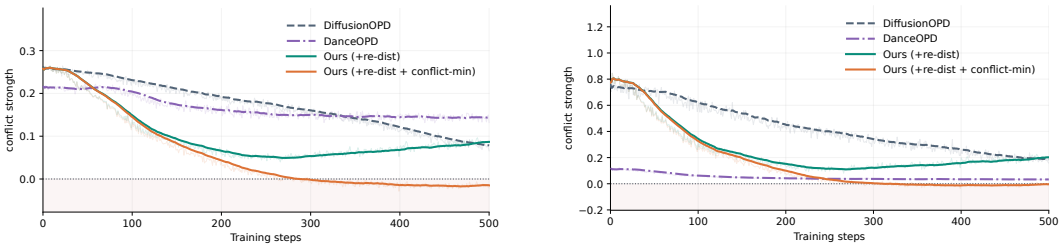
5 Experiments

Our evaluation asks whether the joint optimization transfers multiple teacher skills without conflicts and preserves student capabilities. We first introduce the general setup in Sec. 5.1

5.1 Setup

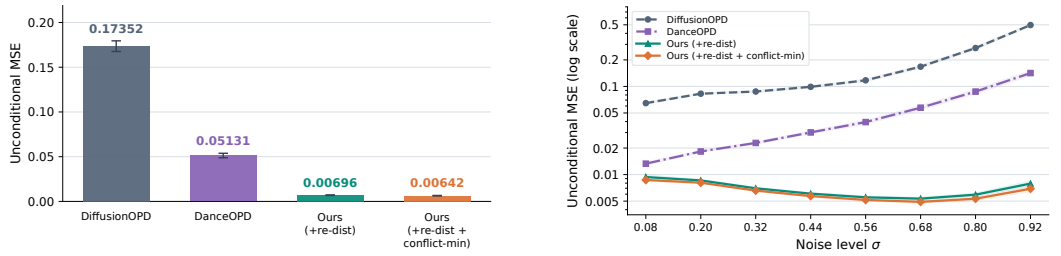
Models and training. We use SD3.5-Medium as the student model (Esser et al., 2024; Stability AI, 2024) and train three RL teachers specializing in aligning human preference (PickScore), OCR, and GenEval for MOPD. Different from the setting of Li et al. (2026) who obtained the teachers with varied RL algorithms (Liu et al., 2025; Zheng et al., 2025) and classifier-free guidance (CFG) scales, we unify the RL training by using diffusionNFT (Zheng et al., 2025) without activating CFG and fix the distillation training with $\text{CFG} = 4$.

Evaluation. We evaluate in-domain capabilities including PickScore (Kirstain et al., 2023), OCR accuracy, and GenEval (Ghosh et al., 2023). We additionally evaluate out-of-domain capabilities



(a) Velocity Conflict (b) Gradient Conflict

Figure 5: **Velocity/Gradient Conflict over Training time.** The velocity conflict (left, measured by cosine similarity) and gradient conflict (right, measured by amplitude) on DiffusionOPD, DanceOPD and GOJO with re-distribution minimization and mutual conflict minimization.



(a) Overall Model Drift (b) Model Drift across Noise level

Figure 6: **GOJO preserves student local geometry.** Comparisons of unconditional generation mean-squared error before and after OPD, where sub-figure (a) shows the results averaged over prompts and denoising steps, and sub-figure (b) shows the results across noise levels. Both suggest a significant drop in the model prior drift.

outside these three objectives, including text–image alignment and human-preference proxies such as CLIPScore (Hessel et al., 2021), HPSv2 (Wu et al., 2023), and ImageReward (Xu et al., 2023).

5.2 Main Results and Behaviour Analysis

Main Comparison. Tab. 1 summarizes the MOPD performance across both in-domain metrics and out-of-domain metrics. It is clear that GOJO outperforms other OPD methods significantly in GenEval, OCR, and PickScore. In other out-of-domain metrics, GOJO also preserves the highest average score even though it slightly underperforms DanceOPD in ClipScore and ImgRwd. We then evaluate whether GOJO truly mitigates velocity conflicts and model prior drift.

Velocity Conflict Reduction. Fig. 5 measure the velocity and gradient conflicts as defined in Sec. 3. It is clear that in the late training stage, both velocity and gradient conflicts are significantly reduced.

Unconditional drift and retention. We use held-out unconditional velocity drift $\mathbb{E}\left[\|v_\theta(x_t, t, \varnothing) - v_{\text{base}}(x_t, t, \varnothing)\|^2\right]$ to measure student model retention. Its connection to KL divergence of two models can be found in Appendix D.3. Figure 6 compares the unconditional drift for DiffusionOPD, DanceOPD, and our method, where we find that our method significantly avoid model prior drifts.

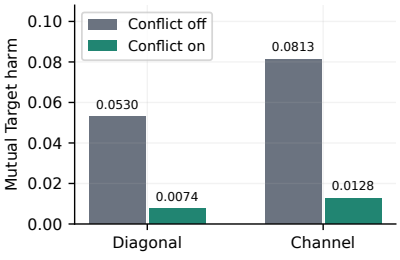


Figure 4: Mutual-teacher protection

5.3 Ablations

Component Analysis. We analyze the impacts of each component, specifically the student geometry preservation (λ), the velocity conflict resolving (γ),

Method / Route	Covariance	γ	λ	GenEval ↑	OCR ↑	PickScore ↑	H_+ ↓	RMS
Soft-Route	–	0	0	0.671	0.703	0.871	–	–
Hard-Route	–	0	0	0.801	0.786	0.894	–	–
Hard-Route	Diagonal	0	1	0.873	0.948	0.912	0.007	0.278
Hard-Route	Channel	0	1	0.878	0.950	0.909	0.013	0.352
Soft-Route	Diagonal	1	1	0.729	0.874	0.880	0.000	0.193
Soft-Route	Channel	1	1	0.744	0.867	0.880	0.000	0.268
Hard-Route	Isotropic	1	1	0.733	0.872	0.880	-	0.180
Hard-Route	Diagonal	1	1	0.880	0.956	0.916	0.053	0.360
Hard-Route	Channel	1	1	0.880	0.954	0.915	0.081	0.464

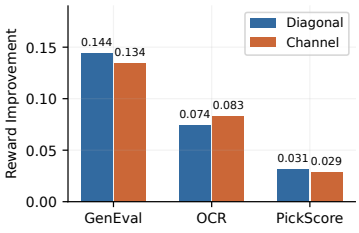


Figure 7: Hard routing and Covariance Anslysis

Table 2: **Component analysis on SD3.5-M.** Ablations of covariance, student geometry (λ), and teacher protection (γ). Rewards average 11 monitoring evaluations at steps 800–1000.

and the covariance type. See Appendix F for a detailed experimental setting. Tab. 2 summarizes the results. As shown in Fig. 7, holding geometry and conflict ($\gamma = 1$) fixed, hard routing improves all tasks GenEval/OCR/PickScore with both diagonal and channel covariance. The hyperparameter settings can be found in Appendix F.

Mutual teacher conflict. With all other fixed, we measure the mutual teacher conflict via:

$$H_+ = \mathbb{E} \sum_i q_i \left[\|v_p - v_{T_i}\|^2 - \|v_{\text{old}} - v_{T_i}\|^2 \right]_+ \tag{15}$$

The results in Fig. 4 shows reductions of 86.0% and 84.3% for diagonal and channel covariance respectively. However, we wish to point out that although teacher-conflict protection individually improves model perofrmance in Tab. 2; it does not further improve the score significantly together with student geometry preservation.

Student geometry preservation. As shown in Fig. 7, student geometry alone already provides a reward boost on GenEval/OCR/PickScoreboost of $8.9\%/20.6\%/2.0\%$ for hard-routing and $8.6\%/11.2\%/-2.0\%$ for soft-routing. We thus conclude that student geometry preservation can significantly improve the OPD performance.

Covariance structure and update magnitude. We compare full channel covariance, diagonal channel covariance in Tab. 2, and isotropic covariance. It is shown that both diagonal- and channel-parameterized covariance can improve the model performance, while the performance between diagonal and channel coavariance does not differ significantly.

6 Discussion and Limitations

GOJO views multi-teacher distillation as a joint optimization problem over teacher agreement, local probability redistribution, and interference between teachers. This perspective provides an interpretable way to construct distillation targets while retaining a simple velocity-regression training objective. However, utilizing up to second momentum to estiamte the velocity geometry can be limited, and we rely on a approximated covariance to estimate the student's local geometry. Furthermore, GOJO relies on the assumption that renoising can be used as a surrogate for on-policy rollout states, which we empirically validate in the appendix but can be broken if the forward noise schedule is changed.

AI Use Statement

Generative AI tools were used to assist literature lookup, mathematical exposition, and manuscript editing.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2): 251–276, 1998. doi: 10.1162/089976698300017746. URL https://doi.org/10.1162/089976698300017746.

Andreas Bergmeister, Stefanie Jegelka, Nikolas Nüsken, Carles Domingo-Enrich, and Jakiw Pidstrigach. Reinforce adjoint matching: Scaling rl post-training of diffusion and flow-matching models, 2026. URL https://arxiv.org/abs/2605.10759.

Tony F. Chan, Gene H. Golub, and Randall J. LeVeque. Algorithms for computing the sample variance: Analysis and recommendations. *The American Statistician*, 37(3):242–247, 1983. doi: 10.1080/00031305.1983.10483115. URL https://doi.org/10.1080/00031305.1983.10483115.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Shiyuan Feng, Huan-ang Gao, Haohan Chi, Hanlin Wu, Zhilong Zhang, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Weak-to-strong generalization via direct on-policy distillation. *arXiv preprint arXiv:2607.05394*, 2026. URL https://arxiv.org/abs/2607.05394.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Abstract-Datasets_and_Benchmarks.html.

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. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/73aacd8b3b05b4b503d58310b523553c-Abstract-Conference.html.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Jeongjae Lee, Jinho Chang, Jeongsol Kim, and Jong Chul Ye. Reward score matching: Unifying reward-based fine-tuning for flow and diffusion models, 2026. URL https://arxiv.org/abs/2604.17415.

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*, 2026. URL https://arxiv.org/abs/2605.15055.

Mingfeng Lin, Chengfei Cai, Lin Xu, Yuxiang Wei, and Liang Han. Dreopd: Degraded-reference extrapolative on-policy distillation for flow-matching models, 2026. URL https://arxiv.org/abs/2608.09233.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.nips.cc/paper/2021/hash/9d27fdf2477ffbff837d73ef7ae23db9-Abstract.html.

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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=oCBKGw5HNf.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.03003.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, Mojie Chi, Xuyan Chi, Jian Cong, Qinpeng Cui, Fei Ding, Qide Dong, Yujiao Du, Haojie Duanmu, Junliang Fan, Jiarui Fang, Jing Fang, Zetao Fang, Chengjian Feng, Yu Gao, Diandian Gu, Dong Guo, Hanzhong Guo, Qiushan Guo, Boyang Hao, Hongxiang Hao, Haoxun He, Jiaao He, Qian He, Tuyen Hoang, Heng Hu, Ruoqing Hu, Yuxiang Hu, Jiancheng Huang, Weilin Huang, Zhaoyang Huang, Zhongyi Huang, Jishuo Jin, Ming Jing, Ashley Kim, Shanshan Lao, Yichong Leng, Bingchuan Li, Gen Li, Haifeng Li, Huixia Li, Jiashi Li, Ming Li, Xiaojie Li, Xingxing Li, Yameng Li, Yiying Li, Yu Li, Yueyan Li, Chao Liang, Han Liang, Jianzhong Liang, Ying Liang, Wang Liao, J. H. Lien, Shanchuan Lin, Xi Lin, Feng Ling, Yue Ling, Fangfang Liu, Jiawei Liu, Jihao Liu, Jingtuo Liu, Shu Liu, Sichao Liu, Wei Liu, Xue Liu, Zuxi Liu, Ruijie Lu, Lecheng Lyu, Jingting Ma, Tianxiang Ma, Xiaonan Nie, Jingzhe Ning, Junjie Pan, Xitong Pan, Ronggui Peng, Xueqiong Qu, Yuxi Ren, Yuchen Shen, Guang Shi, Lei Shi, Yinglong Song, Fan Sun, Li Sun, Renfei Sun, Wenjing Tang, Boyang Tao, Zirui Tao, Dongliang Wang, Feng Wang, Hulin Wang, Ke Wang, Qingyi Wang, Rui Wang, Shuai Wang, Shulei Wang, Weichen Wang, Xuanda Wang, Yanhui Wang, Yue Wang, Yuping Wang, Yuxuan Wang, Zijie Wang, Ziyu Wang, Guoqiang Wei, Meng Wei, Di Wu, Guohong Wu, Hanjie Wu, Huachao Wu, Jian Wu, Jie Wu, Ruolan Wu, Shaojin Wu, Xiaohu Wu, Xinglong Wu, Yonghui Wu, Ruiqi Xia, Xin Xia, Xuefeng Xiao, Shuang Xu, Bangbang Yang, Jiaqi Yang, Runkai Yang, Tao Yang, Yihang Yang, Zhixian Yang, Ziyan Yang, Fulong Ye, Bingqian Yi, Xing Yin, Yongbin You, Linxiao Yuan, Weihong Zeng, Xuejiao Zeng, Yan Zeng, Siyu Zhai, Zhonghua Zhai, Bowen Zhang, Chenlin Zhang, Heng Zhang, Jun Zhang, Manlin Zhang, Peiyuan Zhang, Shuo Zhang, Xiaohe Zhang, Xiaoying Zhang, Xinyan Zhang, Xinyi Zhang, Yichi Zhang, Zixiang Zhang, Haiyu Zhao, Huating Zhao, Liming Zhao, Yian Zhao, Guangcong Zheng, Jianbin Zheng, Xiaozheng Zheng, Zerong Zheng, Kuan Zhu, and Feilong Zuo. Seedance 2.0: Advancing video generation for world complexity, 2026. URL https://arxiv.org/abs/2604.14148.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Stability AI. Stable Diffusion 3.5 Medium model card. Hugging Face model card, 2024. URL https://huggingface.co/stabilityai/stable-diffusion-3.5-medium. Accessed September 12, 2026.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Trans. Mach. Learn. Res.*, 2024, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

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. URL https://arxiv.org/abs/2306.09341.

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 *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/33646ef0ed554145eab65f6250fab0c9-Abstract-Conference.html.

Yixian Xu, Kaiyuan Gao, Yuxiang Chen, Yilei Chen, Zecheng Tang, Zihao Liu, Zikai Zhou, Deqing Li, Hao Meng, Kuan Cao, Jiahao Li, Jie Zhang, Liang Peng, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiaoyue Chen, Yan Shu, Yanran Zhang, Yi Wang, Yu Wu, Yujia Wu, Zekai Zhang, Zhendong Wang, Xiao Xu, Kun Yan, and Chenfei Wu. Qwen-image-2.0-rl technical report, 2026a. URL https://arxiv.org/abs/2606.27608.

Yixian Xu, Yuanrui Zhang, Shengjie Luo, Liwei Wang, and Di He. Designing reinforcement learning for diffusion models: A unified path-space view, 2026b. URL https://arxiv.org/abs/2608.14430.

Fangxu Yu, Weijia Xu, Michael Xu, Tianyi Zhou, and Zinan Lin. Weak-to-strong on-policy distillation. *arXiv preprint arXiv:2607.26246*, 2026. URL https://arxiv.org/abs/2607.26246.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=4OsgYD7em5.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, Kun Yan, Liang Peng, Lihan Jiang, Niantong Li, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiao Xu, Xiaoyue Chen, Xihua Wang, Yan Shu, Yanran Zhang, Yi Wang, Yilei Chen, Ying Ba, Yixian Xu, Yujia Wu, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhendong Wang, Zihao Liu, Zikai Zhou, An Yang, Chen Cheng, Chenxu Lv, Dayiheng Liu, Fan Zhou, Hantian Xiong, Hongzhu Shi, Hu Wei, Huihong Zhao, Ivy Liu, Jianwei Zhang, Jiawei Zhang, Kai Chen, Kang He, Levon Xue, Lin Qu, Linhan Tang, Luwen Feng, Minggang Wu, Minmin Sun,

Na Ni, Rui Men, Shuai Bai, Sishou Zheng, Tao Lan, Tianqi Zhang, Tingkun Wen, Wei Wang, Weixu Qiao, Weiyi Lu, Wenmeng Zhou, Xiaodong Deng, Xiaoxiao Xu, Xinlei Fang, Xionghui Chen, Yanan Wang, Yang Fan, Yichang Zhang, Yixuan Xu, Yu Wu, Zhiyuan Ma, and Zhizhi Cai. Qwen-Image-2.0 technical report, 2026. URL https://arxiv.org/abs/2605.10730.

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. URL https://arxiv.org/abs/2509.16117.

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

Kejian Zhu, Zhuoran Jin, Shangqing Tu, Hongbang Yuan, Yushi Bai, Kang Liu, Juanzi Li, and Jun Zhao. SFT conflicts, RL coexists: A theoretical and empirical analysis of multi-task learning for LLMs. *arXiv preprint arXiv:2608.03573*, 2026a. URL https://arxiv.org/abs/2608.03573.

Yuanzhi Zhu, Xi Wang, Stéphane Lathuilière, and Vicky Kalogeiton. Diffusion reinforcement learning via centered reward distillation, 2026b. URL https://arxiv.org/abs/2603.14128.

A Related Work

Distilling diffusion capabilities. In post-training of diffusion models, current successful methods commonly rely on a velocity regression task, such as Flow matching and rectified (Lipman et al., 2023; Liu et al., 2023). SFT is the most straightforward method in in post-training, which simply relies on a flow matching objective to match the model to a data-induced distribution (Wu et al., 2025; Zhao et al., 2026; Xu et al., 2026a). After SFT, RL is widely adaptad to further shift the distribution towards specific capabilities. With the success of GRPO (Shao et al., 2024), FlowGRPO (Liu et al., 2025) was first developed to make GRPO in flow matching possible. However, FlowGRPO is still based on a reward maximization objective. Later, DiffusionNFT uses reward-weighted forward-process regression for diffusion reinforcement learning (Zheng et al., 2025), which utilizes the velocity regression objective and significantly enhances the stability and convergence speed of diffusion RL. Following methods include Reward Score Matching (Lee et al., 2026), centered reward distillation (Zhu et al., 2026b), refinforceadjoint matching (Bergmeister et al., 2026) and KDE (Xu et al., 2026b) all utilize such a flow matching objective. Compared to RL, the literature for OPD is much less. Two notable variants are diffusionOPD and DanceOPD. DiffusionOPD transfers independently trained specialists through velocity matching on student-induced trajectories (Li et al., 2026), and requires the model to mimic the whole trajectory. DanceOPD treats capability sources as velocity fields and routes student-state queries to the corresponding teachers (Zhou et al., 2026), while DreOPD constructs extrapolated teacher supervision using a degraded reference (Lin et al., 2026). When training, DanceOPD relies on a hard-routing mechanism where each epoch only optimizes towards a single teacher, which means they also sample prompts within a same capabilities. On the other hand, diffusionOPD is a soft-routing mechanism where each epoch has multiple teachers and prompts from different tasks.

Weak-to-strong OPD. We wish to introduce a special class of OPD in LLMs, weak-to-strong OPD. W2S-OPD constructs an LLM proxy teacher by adding the logit difference of positive and negative weak models to a stronger base model, then distills it using reverse KL (Yu et al., 2026; Feng et al., 2026). This motivates transferring a capability increment instead of the full model output. Notably, our old student can be replaced by any model, such that the velocity-space increment matching follows this idea, with diffusion regression replacing the token-distribution objective. We wish to point out that a valuable future work can be integrating gradient conflict solving methods into the framework.

Tackling Multi-task conflict. In meta-learning and multi-task learning, solving mutual-task conflicts has been a long-standing problem, and many works have been developed. For instance, PCGrad (Yu et al., 2020) fixes a target optimization objective and projects other conflicting parameter gradients to the axis induced by the target objective, so that they can be cancelled out. On the other hand, CAGrad combines average-loss optimization with protection of the worst local task improvement (Liu et al., 2021). GOJO instead optimizes the proposal velocity at each query using the finite change in quadratic teacher-matching loss. This avoids teacher-specific backward passes, but protection at a query does not guarantee that every task improves after updating shared parameters.

Local probability geometry. Natural-gradient methods account for information geometry in parameter updates, and trust-region policy optimization limits policy changes (Amari, 1998; Schulman et al., 2015). In GOJO, the metric for measuring local probability geometry is the second moment of the minimum credit needed to realize a velocity correction by reweighting. The ideal covariance is conditional on the current query.

B Ordinary Velocity Regression Can Interfere

We derive the shared-parameter effect in Sec. 3 to explain GOJO's motivation. We wish to state that this analysis is based on the assumption that the model is well-trained, i.e., no significant underfitting.

B.1 Advantage centering can remove shared component in RL

We restate the SFT and RL optimization here following Sec. 3. Specifically, when consolidating a specific task capability i, Zhu et al. (2026a) decompose the parameter update into two parts: $g_i := \nabla_\theta \log \pi_\theta(x_i) = \mu + \epsilon_i$ with μ the task-irrelevant part and ϵ_i the task-relevant part w.r.t. task i.

We first show that the SFT optimization retains the shared component and the RL update cancels out the task-irerelevant part μ.

$$g_i^{\mathrm{SFT}} = \mu + \bar{\epsilon}_i, \qquad g_j^{\mathrm{RL}} = \frac{1}{G}\sum_{k=1}^{G} \widehat{A}_{j,k}\epsilon_{j,k}, \qquad \sum_{k=1}^{G} \widehat{A}_{j,k} = 0. \tag{16}$$

Here i and j denote different tasks, and $\bar{\epsilon}_i = G^{-1}\sum_k \epsilon_{i,k}$ with G the group size and $\epsilon_{i,k}$ the k-th sample.

We will show that GRPO-like centered advantages (Shao et al., 2024) can remove the former,

Proof. Consider a group of G samples from task j with parameter scores $g_{j,k} = \mu + \epsilon_{j,k}$, where μ is shared across all samples in the group. Since μ does not depend on the sample index k, it factors out of the advantage-weighted sum:

$$\begin{aligned} g_j^{\mathrm{RL}} &= \frac{1}{G}\sum_{k=1}^{G} \widehat{A}_{j,k}(\mu + \epsilon_{j,k}) \\ &= \underbrace{\frac{\mu}{G}\sum_{k=1}^{G} \widehat{A}_{j,k}}_{=0} + \frac{1}{G}\sum_{k=1}^{G} \widehat{A}_{j,k}\epsilon_{j,k} \\ &= \frac{1}{G}\sum_{k=1}^{G} \widehat{A}_{j,k}\epsilon_{j,k}. \end{aligned}$$

The shared term vanishes because the centered advantages sum to zero. The sample-dependent terms $\epsilon_{j,k}$ generally remain. For SFT on task i, equal weighting instead gives

$$g_i^{\mathrm{SFT}} = \frac{1}{G}\sum_{k=1}^{G}(\mu + \epsilon_{i,k}) = \mu + \bar{\epsilon}_i, \qquad \bar{\epsilon}_i = \frac{1}{G}\sum_{k=1}^{G}\epsilon_{i,k}.$$

Thus, the SFT average retains the shared component. □

Advantage centering therefore removes the response-independent additive term (Zhu et al., 2026a). It does not imply that the remaining update is orthogonal to μ, or that every shared direction is harmful.

B.1.1 How the two components interfere in OPD.

Ideally, the OPD should only distill the capability-relevant part ϵ_i. However, we show that a velocity regression objective would make the shared component μ enter the RL-teacher update. At θ_{old}, fix their query distributions ν_i, ν_j, teacher fields v_{T_i} and v_{T_j}, and weight $w(t)$. For simplicity, we suppress t, c in the following equations.

We assume both teachers are obtained by local updates from θ_{old} in the same parameterization, and Δ_i, Δ_j are the teacher parameter increments. We consider a, b as their step sizes (this depends on the regression weights w_i and w_j).

$$\Delta_i = a(\mu + \bar{\epsilon}_i), \qquad \Delta_j = \frac{b}{G}\sum_{k} \widehat{A}_{j,k}\epsilon_{j,k}, \tag{17}$$

For one step of gradient descent, the SFT correction is thus $d_i(x_t) = J(x_t)\Delta_i$. Substituting this correction into the additional student update gives

$$h\rho\mathbb{E}_{\nu_i}[w(t)J(x_t)^\top d_i(x_t)] = ha\rho H_i(\mu + \bar{\epsilon}_i).$$

Both components therefore enter the student update through H_i. To isolate μ, compare otherwise identical OPD steps with and without this component in the SFT teacher, holding $\bar{\epsilon}_i$ and the RL teacher fixed. Their velocity difference at the RL query is

$$\Delta v_{j,\mu} = ha\rho J(x_t^{(j)}) H_i \mu + O(h^2). \tag{18}$$

This recovers Eq. (8), where we absorb the teacher-step and route factors into $\eta = ha\rho$. It shows how a shared component retained by the SFT teacher can affect the student's RL route, even though the RL teacher's own centered update cancels that component.

B.2 How GOJO Can Attenuate Cross-Teacher Interference

We next analyze whether GOJO can mitigate the cross-teacher interference. Specifically, GOJO adjusts the teacher correction d_i passed to the student. For source task i, it replaces d_i with the optimized correction $d_i' = v_{\mathrm{p},i} - v_{\mathrm{old}}$. Under the same local linearization as Eq. (18), the contribution of this route to the velocity change at an RL query becomes

$$\Delta v_j^{\mathrm{GOJO}} = h\rho \mathbb{E}_{\nu_i}[w(t) J(x_t^{(j)}) J(x_t^{(i)})^\top d_i'(x_t^{(i)})] + O(h^2).$$

Thus, GOJO adjusts the correction transmitted through the shared parameters. To better understand the correction d_i', we refer to the probability redistribution cost that limits this correction according to the student's local covariance Σ. Under source matching with $\lambda > 0$, zero displacement is feasible and causes no teacher damage. Since all objective terms are nonnegative, comparing this choice with the optimum gives

$$\lambda (d_i')^\top \Sigma^\dagger d_i' \le \|d_i\|^2.$$

This is because the same correction magnitude incurs a higher cost along covariance eigenvectors with small positive eigenvalues. GOJO therefore discourages large updates in directions where the student's sampled velocities vary little.

The mutual teacher damage term further penalizes proposals that increase another teacher's matching error relative to the old student at the same query. It therefore discourages improving SFT-teacher agreement at the expense of RL-teacher agreement with larger γ.

Together, these terms balance teacher agreement against changes to the student's local distribution and damage to other teachers. GOJO does not need to explicitly separate μ from the capability-relevant component. *Consequently, these local controls can attenuate interference, but do not guarantee removal of μ or smaller drift at every query.*

C More Theoretical Results related to the Methodology

C.1 Local Probability Geometry and the Joint Target

We provide more theoretical results in addition to Sec. 4. We first derive the probability cost in Appendix C.1.1. We then show how we can transfer the optimization problem over the scalar credit into optimizing a proposal velocity in Appendix C.1.2. Finally, we solve the target problem in Appendix C.1.3.

In this appendix, we fix (x_t, t, c) and suppress t, c throughout. Within each optimization step, the teacher fields and old student are frozen. The old student will be updated at the end of the epoch. We omit the expectations condition and, by default, consider the old-student distribution without other specification. We retain $\delta_A = v_{\mathrm{p}} - v_{\mathrm{old}}$ from the main text, with matching weights $w_i \ge 0$ and $\sum_i w_i = 1$.

C.1.1 Local Geometry Preserving

We prove the following

Proposition 4. *Assuming that x_0 is generated from the frozen student and we independently draw $x_1 \sim \mathcal{N}(0, I)$, and set $x_t = (1-t)x_0 + tx_1$ and $u_t = x_1 - x_0$. For $0 < t < 1$, we follow the main text and use the moments as*

$$v_{\text{old}} = \mathbb{E}[u_t \mid x_t], \qquad \Sigma = \mathbb{E}[(u_t - v_{\text{old}})(u_t - v_{\text{old}})^\top \mid x_t]. \tag{19}$$

where Σ need to be estimated, which we defer to Appendix E.3 for its practical approximation. Then, let $A(u_t)$ be a centered, square-integrable scalar credit under $\pi(u_t \mid x_t)$. For $1 + A \geq 0$, the reweighted distribution $\pi_A = \pi(1 + A)$ has the following properties

$$\begin{aligned}
\mathbb{E}_{\pi_A}[u_t] - v_{\text{old}} &= \mathbb{E}[A(u_t)(u_t - v_{\text{old}}) \mid x_t], \\
\chi^2(\pi_A \| \pi) &= \mathbb{E}[A^2 \mid x_t], \\
D_{\text{KL}}(\pi_A \| \pi) &= \mathbb{E}[A^2 \mid x_t] + O\left(A^3\right).
\end{aligned}$$

Proof. For the KL divergence, the density ratio $\pi_A/\pi = 1 + A$ gives

$$\begin{aligned}
D_{\text{KL}}(\pi_A \| \pi) &= \mathbb{E}_{\pi_A}\left[\log \frac{\pi_A}{\pi}\right] \\
&= \mathbb{E}_{\pi}\left[\frac{\pi_A}{\pi} \log \frac{\pi_A}{\pi} \mid x_t\right] \\
&= \mathbb{E}[(1 + A) \log(1 + A) \mid x_t].
\end{aligned}$$

For fixed x_t, let $\|A\|_\infty$ denote the essential supremum of $|A|$ under $\pi(\cdot \mid x_t)$. When $\|A\|_\infty \leq 1/2$, Taylor expansion gives, uniformly over the sample velocities,

$$\log(1 + A) = A - A^2 + O(|A|^3), \qquad (1 + A) \log(1 + A) = A + A^2 + O(|A|^3).$$

Taking conditional expectations removes the linear term because $\mathbb{E}[A \mid x_t] = 0$. Moreover,

$$\mathbb{E}[|A|^3 \mid x_t] \leq \|A\|_\infty \mathbb{E}[A^2 \mid x_t].$$

Consequently, as $\|A\|_\infty \to 0$,

$$\begin{aligned}
D_{\text{KL}}(\pi_A \| \pi) &= \mathbb{E}[A^2 \mid x_t] + O\big(\mathbb{E}[|A|^3 \mid x_t]\big) \\
&= \mathbb{E}[A^2 \mid x_t] + o\big(\mathbb{E}[A^2 \mid x_t]\big).
\end{aligned}$$

□

C.1.2 From Credit Assignment to Joint Optimization

With the above derivation, we can derive the joint optimization problem as

$$\begin{aligned}
\min_A \sum_k w_k \Big[& \underbrace{P \|v_{\text{old}} + \delta_A - v_{T_k}\|^2}_{\text{Teacher Agreement}} + \lambda \underbrace{\mathbb{E}[A^2 \mid x_0]}_{\text{Local probability cost}} \\
& + \underbrace{\gamma \sum_{j \neq k} q_j \Big[\|v_{\text{old}} + \delta_A - v_{T_j}\|^2 - \|v_{\text{old}} - v_{T_j}\|^2 \Big]_+}_{\text{Mutual Teacher Damage}} \Big]
\end{aligned} \tag{20}$$

The teacher agreement and damage terms depend on the credit function only through the displacement δ_A. Different credit functions can induce the same displacement δ_A and therefore the same teacher losses while having different values of $\mathbb{E}[A^2 \mid x_t]$. Thus, we can first fix δ_A and find the least-cost credit that realizes it. The following proposition solves this inner problem, i.e.,$\min \mathbb{E}[A^2 \mid x_t]$.

Proposition 1 (Minimum local credit equality)**.** For a displacement δ_A, the minimum over credits satisfies

$$\min_{\substack{\mathbb{E}[A|x_t]=0 \\ \mathbb{E}[A(u_t)(u_t - v_{\text{old}})|x_t] = \delta_A}} \mathbb{E}[A^2 \mid x_t] = \delta_A^\top \Sigma^\dagger \delta_A. \tag{21}$$

which is attained by

$$A^*(u_t) = (u_t - v_{\text{old}})^\top \Sigma^\dagger \delta_A.$$

The minimizer is unique up to conditional almost-sure equality.

Proof. We fix x_t and suppress conditioning below.

Step 1: Identify the attainable displacements. For any $e \in \mathrm{Null}(\Sigma)$,

$$\mathbb{E}[(e^\top(u_t - v_{\mathrm{old}}))^2] = e^\top \Sigma e = 0.$$

The squared quantity is nonnegative, so $e^\top(u_t - v_{\mathrm{old}}) = 0$ almost surely. There is no sample-velocity variation in this direction, and reweighting cannot shift the mean along it:

$$e^\top \delta_A = \mathbb{E}[A(u_t) e^\top(u_t - v_{\mathrm{old}})] = 0.$$

Thus every attainable displacement is orthogonal to $\mathrm{Null}(\Sigma)$. Since Σ is symmetric, this is equivalent to $\delta_A \in \mathrm{Range}(\Sigma)$.

Step 2: Construct a credit for a given displacement. Consider a linear credit $A(u_t) = (u_t - v_{\mathrm{old}})^\top b$, where b is a coefficient vector. This credit is centered because $\mathbb{E}[u_t - v_{\mathrm{old}}] = 0$, and its induced displacement is

$$\mathbb{E}[A(u_t)(u_t - v_{\mathrm{old}})] = \mathbb{E}[(u_t - v_{\mathrm{old}})(u_t - v_{\mathrm{old}})^\top] b = \Sigma b.$$

To realize the prescribed δ_A, we therefore solve $\Sigma b = \delta_A$. For $\delta_A \in \mathrm{Range}(\Sigma)$, one solution is $b = \Sigma^\dagger \delta_A$, giving the stated A^*. Its energy is

$$\mathbb{E}[(A^*)^2] = \delta_A^\top \Sigma^\dagger \Sigma \Sigma^\dagger \delta_A = \delta_A^\top \Sigma^\dagger \delta_A.$$

Step 3: We can show that no other feasible credit has lower energy. Let A be any centered, square-integrable credit inducing the same δ_A, and write $r = A - A^*$. Because A and A^* have the same mean and induce the same displacement,

$$\mathbb{E}[r] = 0, \qquad \mathbb{E}[r(u_t - v_{\mathrm{old}})] = \delta_A - \delta_A = 0.$$

Consequently, the cross term in their energy decomposition vanishes:

$$\mathbb{E}[r A^*] = \delta_A^\top \Sigma^\dagger \mathbb{E}[r(u_t - v_{\mathrm{old}})] = 0.$$

Expanding the square now gives

$$\begin{aligned} \mathbb{E}[A^2] &= \mathbb{E}[(A^* + r)^2] \\ &= \mathbb{E}[(A^*)^2] + 2\mathbb{E}[r A^*] + \mathbb{E}[r^2] \\ &= \delta_A^\top \Sigma^\dagger \delta_A + \mathbb{E}[r^2] \geq \delta_A^\top \Sigma^\dagger \delta_A. \end{aligned}$$

Thus any residual credit r that leaves the mean displacement unchanged only adds energy. Equality holds exactly when $\mathbb{E}[r^2] = 0$, or $A = A^*$ almost surely. This proves global optimality over all feasible credits. □

Interestingly, this credit can be interpreted as an implicit credit over a proposal direction$u_t - v_{\mathrm{old}}$. I.e., one proposal to move through to u_t, the teacher direction is $v_{T_k} - v_{\mathrm{old}}$. According to the student's local geometry Σ, the credit measures the agreement of two directions according to the local geometry, which can be expressed (in expectation) as a scalar credit $A^*(u_t) = (u_t - v_{\mathrm{old}})^\top \Sigma^\dagger \delta_A$..

C.1.3 Solving for the Optimal Proposal Velocity

The previous subsection gives the least credit energy needed to realize each displacement. Substituting this value into Eq. equation 10 yields the following equivalent problem, with $\lambda, \gamma \geq 0$.

Proposition 5 (Equivalent velocity objective)**.** *The local credit-assignment problem reduces to*

$$\begin{aligned} \min_{\delta_A \in \mathrm{Range}(\Sigma)} \sum_i w_i \Big[&\|v_{\mathrm{old}} + \delta_A - v_{T_i}\|^2 + \lambda \delta_A^\top \Sigma^\dagger \delta_A \\ &+ \gamma \sum_{j \neq i} q_j \left[\|v_{\mathrm{old}} + \delta_A - v_{T_j}\|^2 - \|v_{\mathrm{old}} - v_{T_j}\|^2 \right]_+ \Big]. \end{aligned} \tag{22}$$

We divide the entire objective by two, which leaves its minimizer unchanged. Accordingly, define half of teacher i's matching-error increase as

$$h_i(v_{\mathrm{p}}) = \|v_{\mathrm{p}} - v_{T_i}\|^2 - \|v_{\mathrm{old}} - v_{T_i}\|^2. \tag{23}$$

Thus $h_i > 0$ means worse agreement with teacher i, while $h_i \le 0$ incurs no damage penalty. Collecting all protection terms for the same teacher gives

$$\eta_i = \gamma q_i \sum_{k \neq i} w_k = \gamma q_i (1 - w_i). \tag{24}$$

The normalized objective is therefore

$$F(v_{\mathrm{p}}) = \sum_i w_i \|v_{\mathrm{p}} - v_{T_i}\|^2 + \frac{\lambda}{2}(v_{\mathrm{p}} - v_{\mathrm{old}})^\top \Sigma^\dagger (v_{\mathrm{p}} - v_{\mathrm{old}}) + \sum_i \eta_i [h_i(v_{\mathrm{p}})]_+. \tag{25}$$

The same derivation also covers protection of all teachers by setting $\eta_i = \gamma q_i$, as in Appendix E.4.

Replacing each positive part with a bounded coefficient. First suppose $\Sigma \succ 0$ such that $\Sigma^\dagger = \Sigma^{-1}$ and every displacement is attainable. For fixed v_{p}, the value $h_i(v_{\mathrm{p}})$ is a scalar, and

$$\eta_i [h_i(v_{\mathrm{p}})]_+ = \max_{0 \le \alpha_i \le \eta_i} \alpha_i h_i(v_{\mathrm{p}}). \tag{26}$$

If $h_i > 0$, the maximum is attained at $\alpha_i = \eta_i$; if $h_i < 0$, it is attained at $\alpha_i = 0$. When $h_i = 0$, every allowed coefficient gives zero. This identity lets us write $F(v_{\mathrm{p}}) = \max_\alpha L(v_{\mathrm{p}}, \alpha)$ over the stated coefficient bounds, where

$$L(v_{\mathrm{p}}, \alpha) = \sum_i (w_i + \alpha_i)\|v_{\mathrm{p}} - v_{T_i}\|^2 + \frac{\lambda}{2}(v_{\mathrm{p}} - v_{\mathrm{old}})^\top \Sigma^{-1}(v_{\mathrm{p}} - v_{\mathrm{old}}) - \sum_i \alpha_i \|v_{\mathrm{old}} - v_{T_i}\|^2. \tag{27}$$

For fixed coefficients, α_i acts as an additional matching weight for teacher i. The last term is constant with respect to v_{p}.

Solving for the velocity when the coefficients are fixed. Setting the gradient of L to zero gives

$$\sum_i (w_i + \alpha_i)(v_{\mathrm{p}} - v_{T_i}) + \lambda \Sigma^{-1}(v_{\mathrm{p}} - v_{\mathrm{old}}) = 0.$$

Since $\sum_i w_i = 1$, the coefficient matrix, also the Hessian of L, is

$$M_\alpha = \left(1 + \sum_i \alpha_i\right) I + \lambda \Sigma^{-1}. \tag{28}$$

It is positive definite, so L has positive curvature in every direction.

The zero-gradient solution is its unique minimum:

$$\begin{aligned} M_\alpha (v_{\mathrm{p}} - v_{\mathrm{old}}) &= \sum_i (w_i + \alpha_i)(v_{T_i} - v_{\mathrm{old}}), \\ v_{\mathrm{p}}(\alpha) &= v_{\mathrm{old}} + M_\alpha^{-1} \sum_i (w_i + \alpha_i)(v_{T_i} - v_{\mathrm{old}}). \end{aligned} \tag{29}$$

Thus the velocity can be computed directly once the coefficients are known. If Σ has zero eigenvalues, the displacement must remain in $\mathrm{Range}(\Sigma)$. For $\lambda > 0$, solving the same velocity equation on this subspace gives

$$v_{\mathrm{p}}(\alpha) = v_{\mathrm{old}} + \Sigma \left[\left(1 + \sum_i \alpha_i\right)\Sigma + \lambda I\right]^{-1} \sum_i (w_i + \alpha_i)(v_{T_i} - v_{\mathrm{old}}). \tag{30}$$

Along an eigenvector of Σ with eigenvalue s, a teacher correction is multiplied by $s/[(1+\sum_i \alpha_i)s + \lambda]$. This factor is zero when $s = 0$, so the proposal cannot move in an unattainable direction. The coefficient optimization and gap remain valid on this subspace, with Σ^{-1} replaced by $\Sigma^\dagger$. At α^*, the expression above gives Eq. (12). The geometry-off variant below sets $\lambda = 0$ and removes the range constraint.

Choosing the coefficients. For any allowed α, $L(v_{\mathrm{p}}, \alpha) \le F(v_{\mathrm{p}})$. Its minimum is therefore a lower bound on the minimum of F. Substituting the velocity above, or completing the square, gives

$$\begin{aligned} \min_{v_{\mathrm{p}}} L(v_{\mathrm{p}}, \alpha) &= \sum_i w_i \|v_{\mathrm{old}} - v_{T_i}\|^2 - B(\alpha), \\ B(\alpha) &= (v_{\mathrm{p}}(\alpha) - v_{\mathrm{old}})^\top M_\alpha (v_{\mathrm{p}}(\alpha) - v_{\mathrm{old}}), \\ \alpha^* &\in \arg\min_{0 \le \alpha_i \le \eta_i} B(\alpha). \end{aligned} \tag{31}$$

The first term is constant, so minimizing B selects the largest lower bound. This is the dual problem: it has one bounded scalar variable per teacher.

To see why this also recovers the original optimum, differentiate the velocity equation to obtain $\partial_i v_{\mathrm{p}} = M_\alpha^{-1}(v_{T_i} - v_{\mathrm{p}})$. Then

$$\begin{aligned} \partial_i B &= (v_{T_i} - v_{\mathrm{old}})^\top (v_{\mathrm{p}} - v_{\mathrm{old}}) - \|v_{\mathrm{p}} - v_{\mathrm{old}}\|^2 = -h_i(v_{\mathrm{p}}), \\ \partial^2_{ij} B &= (v_{T_i} - v_{\mathrm{p}})^\top M_\alpha^{-1} (v_{T_j} - v_{\mathrm{p}}). \end{aligned} \tag{32}$$

Here $v_{\mathrm{p}} = v_{\mathrm{p}}(\alpha)$. The second derivative matrix is positive semidefinite, so B is convex: any local minimum is also a global minimum. For $\eta_i > 0$, if $h_i(v_{\mathrm{p}}(\alpha^*)) > 0$, then $\partial_i B < 0$ and increasing α_i would reduce B unless $\alpha_i^* = \eta_i$. Similarly, negative damage requires $\alpha_i^* = 0$; an interior coefficient requires zero damage. For $\eta_i = 0$, the coefficient is fixed at zero.

These conditions give $\alpha_i^* h_i(v_{\mathrm{p}}(\alpha^*)) = \eta_i [h_i(v_{\mathrm{p}}(\alpha^*))]_+$ for every teacher. Hence L equals F at the resulting proposal, and for any velocity v_{p},

$$F(v_{\mathrm{p}}) \ge L(v_{\mathrm{p}}, \alpha^*) \ge L(v_{\mathrm{p}}(\alpha^*), \alpha^*) = F(v_{\mathrm{p}}(\alpha^*)).$$

This proves that $v_{\mathrm{p}}(\alpha^*)$ solves the original velocity problem. The coefficients are found numerically for each query and the velocity formula is closed form once they are known.

C.1.4 Reductions and Scope of the Guarantees

No damage penalty. With $\eta_i = 0$ and $\lambda > 0$, $\alpha^* = 0$ and

$$v_{\mathrm{p}} = v_{\mathrm{old}} + \Sigma(\Sigma + \lambda I)^{-1} \sum_i w_i (v_{T_i} - v_{\mathrm{old}}). \tag{33}$$

For hard routing with $w_k = 1$, only teacher k appears in the sum; the other tasks still train the same student. If this teacher alone participates in both matching and protection, the solution remains optimal for any damage weight, because stationarity gives

$$h_k(v_{\mathrm{p}}) = -\|v_{\mathrm{p}} - v_{\mathrm{old}}\|^2 - \lambda (v_{\mathrm{p}} - v_{\mathrm{old}})^\top \Sigma^\dagger (v_{\mathrm{p}} - v_{\mathrm{old}}) \le 0. \tag{34}$$

Other teachers can still require protection in multi-teacher training.

No geometry penalty. With geometry disabled,

$$v_{\mathrm{p}} = \frac{\sum_i (w_i + \alpha_i^*) v_{T_i}}{1 + \sum_i \alpha_i^*}, \text{ and } B(\alpha) = \frac{\|\sum_i (w_i + \alpha_i)(v_{T_i} - v_{\mathrm{old}})\|^2}{2(1 + \sum_i \alpha_i)}. \tag{35}$$

The target is a convex combination of teacher fields, with weights determined by damage relative to the old student. If the damage penalty is also disabled, $v_{\mathrm{p}} = \sum_i w_i v_{T_i}$, recovering the direct OPD target for the same queries and effective teacher fields. Appendix D.1 discusses the remaining differences between complete training recipes.

D Theoretical Results for Interpretability

D.1 Connections to Implicit Credit and Transfer Targets

D.1.1 Teacher Corrections as Posterior Credits

When teacher and old-student velocities are exact conditional means under the same forward process, their difference also admits a posterior-credit interpretation. Assume $p_{T_k} \ll p_{\mathrm{old}}$, finite moments,

and a positive finite noisy-marginal density ratio at the query. Bayes' rule gives the centered credit

$$A(u_t) = \frac{p_{T_k}(x_0 \mid x_t)}{p_{\mathrm{old}}(x_0 \mid x_t)} - 1 = \frac{p_{T_k}(x_0)}{p_{\mathrm{old}}(x_0)} \frac{p_{\mathrm{old}}(x_t)}{p_{T_k}(x_t)} - 1, \tag{36}$$

where $x_0 = x_t - tu_t$ at the fixed query. Consequently,

$$\mathbb{E}[A \mid x_t] = 0, \qquad \mathbb{E}[A(u_t)(u_t - v_{\mathrm{old}}) \mid x_t] = v_{T_k} - v_{\mathrm{old}}. \tag{37}$$

The posterior ratio integrates to one, and its product with u_t integrates to the teacher's conditional mean velocity. If this credit is square-integrable, Proposition 1 gives a minimum-energy linear credit with the same correction. That linear credit need not equal the posterior ratio minus one, recover the teacher's historical reward, or satisfy $A \geq -1$. Guided teacher fields remain valid inputs to GOJO without assuming that they are exact conditional-mean fields of an endpoint distribution.

GOJO constructs a velocity target from supplied teacher fields. This section relates the target to scalar credits, teacher guidance, and reference-relative transfer. We use the fixed-query notation and self-consistency assumption of Appendix C.1.1.

Geometry-only GOJO as regularized credit assignment. With no damage penalty, optimizing centered, square-integrable credits gives

$$\min_A \ \frac{1}{2} \left\| \delta_A - \sum_i w_i (v_{T_i} - v_{\mathrm{old}}) \right\|^2 + \frac{\lambda}{2} \mathbb{E}[A^2 \mid x_t], \qquad \lambda > 0, \tag{38}$$

where $\delta_A = \mathbb{E}[A(u_t)(u_t - v_{\mathrm{old}}) \mid x_t]$. Proposition 1 gives

$$\begin{aligned} A^*(u_t) &= (u_t - v_{\mathrm{old}})^\top (\Sigma + \lambda I)^{-1} \sum_i w_i (v_{T_i} - v_{\mathrm{old}}), \\ \delta_{A^*} &= \Sigma (\Sigma + \lambda I)^{-1} \sum_i w_i (v_{T_i} - v_{\mathrm{old}}). \end{aligned} \tag{39}$$

The covariance filter follows from the quadratic credit cost and remains well-defined for singular Σ. The implementation directly constructs the mean displacement and requires no reward network or sampled credit target.

Unregularized stochastic credit targets. For a prescribed attainable proposal v_{p}, the minimum-energy credit $A^*(u_t) = (u_t - v_{\mathrm{old}})^\top \Sigma^\dagger (v_{\mathrm{p}} - v_{\mathrm{old}})$ has mean displacement $v_{\mathrm{p}} - v_{\mathrm{old}}$. For any baseline b fixed at the query, centering gives

$$\mathbb{E}[A^*(u_t)(u_t - b) \mid x_t] = v_{\mathrm{p}} - v_{\mathrm{old}}. \tag{40}$$

With finite target second moments, regression to the detached random target $v_{\mathrm{old}} + A^*(u_t)(u_t - b)$ equals regression to v_{p} plus a parameter-independent conditional variance. Choosing the teacher as the baseline therefore leaves the population optimum unchanged and does not recover its historical RL gradient. Minimizing scalar-credit energy also need not minimize target variance. GOJO changes the mean target through its probability and damage penalties; clipping credits, restricting directions, or using sample-dependent baselines would give different estimators.

Probability ratios and velocity differences. With frozen p_{old} and suitable support and normalization, the reward $r = \beta \log(p_{T_k}/p_{\mathrm{old}})$ makes p_{T_k} the optimizer of $\mathbb{E}_p[r] - \beta D_{\mathrm{KL}}(p \| p_{\mathrm{old}})$. A teacher/reference reward $\log(p_{T_k}/p_{\mathrm{ref}})$ with a student base p_B instead gives

$$p^* \propto p_B (p_{T_k}/p_{\mathrm{ref}})^{1/\beta}. \tag{41}$$

This is reference-relative policy transfer (Feng et al., 2026). A diffusion velocity query does not directly provide endpoint density ratios. The posterior-credit identity in Appendix D.1.1 connects a density ratio to a mean velocity correction under its stated assumptions; an arbitrary velocity difference need not be the marginal field of this endpoint reweighting.

D.2 Connection to previous methods

Weak-to-strong Distillation When the covariance matrix is isotropic and the velocity conflict is off. GOJO would reduce to weak-to-strong distillation (Yu et al., 2026; Lin et al., 2026; Feng et al., 2026). Specifically, the matching can be rewritten as

$$v_{\mathrm{p},i} = v_{\text{base},c} + a_i(v_{T_i,c} - v_{\mathrm{ref},i,c}), \qquad a_i \geq 0. \tag{42}$$

The existing W2S-OPD method operates on language-model logits with reverse-KL distillation (Yu et al., 2026); Eq. (42) describes a velocity-space analogue.

Connection to CFG. When the covariance matrix is isotropic, the base model equals the reference model; our GOJO will be reduced to a simple classifer-free guidance that,

$$\begin{aligned} v_{T_i}^{\mathrm{self-cfg}} &= v_{\text{base},\varnothing} + (v_{T_i,c} - v_{\text{base},\varnothing}), \\ v_{T_i}^{\mathrm{self-cfg}} - v_{\mathrm{old}} &= (v_{\text{base},\varnothing} - v_{\mathrm{old}}) + (v_{T_i,c} - v_{\text{base},\varnothing}). \end{aligned} \tag{43}$$

This is as simple as a classifier-free guidance method.

D.3 Connection to KL under Gaussian transitions

For a local stochastic reference model, suppose student and base transitions share a positive isotropic variance:

$$p_m(x_{t-\Delta t} \mid x_t, \varnothing) = \mathcal{N}\big(x_t - \Delta t\, v_m(x_t, t, \varnothing),\, s_t^2 I\big)\,, \qquad m \in \{\theta, \text{base}\}, \quad s_t > 0. \tag{44}$$

For Gaussians with identical covariance, the covariance and log-determinant terms in the KL cancel. Substituting their mean difference gives

$$D_{\mathrm{KL}}(p_\theta(\cdot \mid x_t, \varnothing) \,\|\, p_{\text{base}}(\cdot \mid x_t, \varnothing)) = \frac{(\Delta t)^2}{2s_t^2} \|v_\theta(x_t, t, \varnothing) - v_{\text{base}}(x_t, t, \varnothing)\|^2. \tag{45}$$

Thus the per-coordinate drift measures the same local quadratic discrepancy, up to the factor $D(\Delta t)^2/(2s_t^2)$. This motivates velocity drift as a retention proxy. Deterministic DPM2 does not use these Gaussian transitions, so the measured drift does not estimate its transition KL or the KL between unconditional image distributions. Averaging over the reference queries also gives no certified endpoint-KL bound.

E Implementation Details

We detail the implementation in Section 4.3, from constructing queries and estimating covariance to solving for the proposal velocity and training the student.

E.1 Query Construction and Classifier-free Guidance

Query Details GOJO generates an endpoint x_0 with the frozen old student, samples an independent $\epsilon \sim \mathcal{N}(0, I)$ and an interior time t, and forms

$$x_t = (1-t)x_0 + t\epsilon, \qquad u_t = \epsilon - x_0. \tag{46}$$

The old student, trainable student, and all teachers are evaluated at the same query (x_t, t, c). Re-noised endpoints generally differ in distribution from intermediate rollout states, even with equal rollout lengths or query counts. The local probability derivation uses this forward process and does not require the two query distributions to coincide.

CFG details The teacher fields in the objective include guidance. We use the frozen base model's unconditional prediction as the anchor with guidance scale 4 (DOG4 in Eq. equation 43). Teacher self-CFG uses each teacher's own unconditional prediction. Component comparisons keep the query distribution and guidance fixed.

E.2 HARD-ROUTING VS SOFT-ROUTING

For a query from source task k, hard routing sets $w_j = \mathbf{1}[j = k]$; soft routing sets $w_j = 1/K$. Across the whole training procedure, both take all tasks but only differ in which teachers enter the matching term at each epoch.

Using the normalized objective in Eq. (25), the implemented all-teacher protection sets $\eta_j = \gamma q_j$ with $q_j = 1/K$ for both routing policies. Thus hard routing uses

$$\frac{1}{2}\|v_{\mathrm{p}} - v_{T_k}\|^2 + \frac{\lambda}{2}(v_{\mathrm{p}} - v_{\mathrm{old}})^\top \Sigma^\dagger (v_{\mathrm{p}} - v_{\mathrm{old}}) + \frac{\gamma}{K}\sum_j [h_j(v_{\mathrm{p}})]_+, \tag{47}$$

Soft routing replaces the first term by $\frac{1}{2K}\sum_j \|v_{\mathrm{p}} - v_{T_j}\|^2$. If the matched teacher is excluded from protection, the coefficients become $\eta_j = \gamma q_j(1 - w_j)$, as in Appendix C.1.3.

Both policies use the covariance filter and dual solver in Appendix C.1.3. Under hard routing, other teachers enter the target through protection, while only the source teacher enters the matching term. Disabling both penalties recovers v_{T_k} for hard routing and $K^{-1}\sum_j v_{T_j}$ for soft routing.

E.3 COVARIANCE MODELS AND ESTIMATION

The implementation uses each spatial residual vector $r_p = (x_1 - x_0 - v_{\mathrm{old}})_p \in \mathbb{R}^C$ as an observation in a source-task/time bank. With B time buckets, the index is $\lfloor Bt \rfloor$ capped at $B - 1$. The channel covariance is centered across queries and spatial positions within the bank.

Let $\widehat{\Sigma}_{\mathrm{ch}}$ be this $C \times C$ estimate, as in the main text. With channel-major flattening, the modeled latent covariances are

$$\widehat{\Sigma} \in \left\{ \widehat{\Sigma}_{\mathrm{ch}} \otimes I_P \text{ (channel)},\ \mathrm{diag}(\boldsymbol{c}) \otimes I_P \text{ (diagonal)},\ pI_{CP} \text{ (isotropic)} \right\}. \tag{48}$$

Channel covariance retains cross-channel correlations, while all three modes omit cross-position covariance. Diagonal covariance allows different channel variances; isotropic covariance shares one variance. Zero covariance alone does not imply independence.

Centering remains necessary after subtracting v_{old}. Under the bank's observation law, total covariance gives

$$\begin{aligned}\mathrm{Cov}(r_p \mid \text{bank}) = {} & \mathbb{E}[\mathrm{Cov}(r_p \mid x_t, t, c, p, \text{bank}) \mid \text{bank}] \\ & + \mathrm{Cov}(\mathbb{E}[r_p \mid x_t, t, c, p, \text{bank}] \mid \text{bank}).\end{aligned} \tag{49}$$

The first term averages local covariance; the second includes variation in the conditional-mean error. Thus, centering the pooled observations does not recover a query-specific Σ. Spatial correlation also means that the observation count may overstate the number of independent samples.

Centered updates and snapshot refresh. Each bank maintains observation count N, channel mean $\bar{r}$, and centered outer-product sum $S = \sum_\ell (r_\ell - \bar{r})(r_\ell - \bar{r})^\top$, plus a complete-query count. These statistics are saved in checkpoints and are not trainable. We use the covariance estimate S/N.

Each global minibatch is aggregated in two passes: count and mean, then outer products centered by the global mean. Every physical rank contributes local observations once; teacher-parallel copies are not gathered and counted again. For previous statistics $(N, \bar{r}, S)$ and new minibatch statistics $(N_b, \bar{r}_b, S_b)$, the centered merge is (Chan et al., 1983)

$$\begin{aligned} N' &= N + N_b, \qquad \bar{r}' = \bar{r} + \frac{N_b}{N'}(\bar{r}_b - \bar{r}), \\ S' &= S + S_b + \frac{N N_b}{N'}(\bar{r}_b - \bar{r})(\bar{r}_b - \bar{r})^\top. \end{aligned} \tag{50}$$

Empty minibatches contribute nothing; an empty bank adopts the nonempty minibatch statistics. This avoids cancellation from subtracting two large raw moments.

A nonempty bank supplies its previous statistics to the current target and merges the minibatch afterward. An empty bank uses the current global minibatch for initialization. Using previous statistics reduces dependence on the current batch, although queries may remain correlated. Banks reset whenever the frozen old student is refreshed, so each bank uses a single anchor.

E.4 Numerical Solver and Computational Cost

For channel covariance, diagonalize each accessed bank once: $\widehat{\Sigma}_{\mathrm{ch}} = Q\,\mathrm{diag}(s_c)Q^\top$. For diagonal and isotropic covariance, use $Q = I$ and the selected channel variances as s_c. For each query, rotate the teacher corrections along channels and cache

$$G_c[i,j] = \frac{1}{CP}\sum_p [Q^\top(v_{T_i} - v_{\mathrm{old}})]_{c,p}[Q^\top(v_{T_j} - v_{\mathrm{old}})]_{c,p}. \tag{51}$$

At each dual iterate, compute the channel gains and weighted Gram matrices

$$r_c = \begin{cases} s_c/[(1+\sum_i \alpha_i)s_c + \lambda], & \lambda > 0, \\ 1/(1+\sum_i \alpha_i), & \lambda = 0, \end{cases} \qquad K_j = \sum_c r_c^j G_c, \quad j = 1,2,3.$$

With geometry disabled, no covariance bank or rotation is needed. Writing $w + \alpha$ for the vector of combined teacher weights, the dual objective in Appendix C.1.3 and its derivatives, normalized by $D = CP$, are

$$\begin{aligned} \frac{B(\alpha)}{D} &= \tfrac{1}{2}(w+\alpha)^\top K_1(w+\alpha), \\ \frac{(\nabla B(\alpha))_i}{D} &= [K_1(w+\alpha)]_i - \tfrac{1}{2}(w+\alpha)^\top K_2(w+\alpha), \\ \frac{(\nabla^2 B(\alpha))_{ij}}{D} &= (K_1)_{ij} - [K_2(w+\alpha)]_i - [K_2(w+\alpha)]_j \\ &\quad + (w+\alpha)^\top K_3(w+\alpha). \end{aligned} \tag{52}$$

After solving for α^*, reconstruct the proposal once using the corresponding gains:

$$[Q^\top(v_{\mathrm{p}} - v_{\mathrm{old}})]_{c,p} = r_c \sum_i (w_i + \alpha_i^*)[Q^\top(v_{T_i} - v_{\mathrm{old}})]_{c,p}. \tag{53}$$

The solver uses bounded damped Newton iterations on free box coordinates, projected-gradient fallback, and feasible backtracking. Working-set refinement handles directions pointing out of active bounds. Accepted steps must pass an objective-decrease check. Defaults are 64 iterations, 12 backtracking scales, absolute gap tolerance 10^{-7}, and relative tolerance 10^{-5} against old matching loss, with mean-normalized objectives. The retained iterate, gap, convergence status, and line-search failures are reported; nonconvergence is not silently replaced by OPD. Newton damping modifies the search direction, not the target objective or covariance.

Gram caching costs $O(CPK^2)$ after channel rotation, which costs $O(KC^2P)$; bank eigendecomposition costs $O(C^3)$. Dual evaluations use the cached matrices and K-dimensional solves. There is no additional teacher forward when the existing executor already supplies every same-query teacher output. Extending a source-only executor to protect additional teachers would increase forward cost.

E.5 Detached Regression and the Training Loop

The student fits the detached proposal at each query with

$$\frac{1}{2}\left\|v_\theta(x_t,t,c) - \mathrm{sg}[v_{\mathrm{p}}(x_t,t,c)]\right\|^2, \tag{54}$$

averaged over old-student re-noised queries with the outer weights described below. Only θ receives gradients. Teachers and the old student remain frozen; covariance moments are maintained as statistics, and the dual coefficients are solved for each query. Credit is eliminated analytically, so no reward network or sampled credit–residual target is needed.

Outer weighting and unconditional consistency. The training recipe also uses time weighting, lagged RMS compensation, and a separate base-unconditional consistency loss. RMS compensation tracks the energy of the matching-weighted correction $\sum_j w_j(v_{T_j} - v_{\mathrm{old}})$ relative to the proposal displacement and uses their square-root ratio as a bounded conditional-loss weight. Its exact-time EMA, warm-up, and ramp-up are separate from the covariance bank and persist across old-student

Algorithm 1 GOJO training with pooled probability geometry

Require: Student θ; frozen teachers; task and time samplers; λ; matching weights w; protection coefficients η; outer weighting

for each outer block **do**
  Freeze the old student and clear covariance banks
  Generate endpoints x_0 on sampled source-task prompts
  for each minibatch **do**
    Sample t, ϵ and form x_t and u_t by Eq. (46)
    Evaluate detached v_{old} and effective teacher fields at (x_t, t, c)
    Select channel, diagonal, or isotropic geometry.
    Solve $0 \leq \alpha_j \leq \eta_j$
    Reconstruct and detach v_{p}
    Calculate the student regression loss and any outer consistency loss
    Update θ at the gradient-accumulation boundary
  end for
end for

refreshes. This factor changes the regression weight, not the target or the unconditional loss. The unconditional consistency term is also an outer training choice, separate from the three inner-objective terms. A positive scalar weight preserves the pointwise regression target but affects training through shared parameters. Target ablations therefore hold guidance, query sampling, matching, protection eligibility, and the outer recipe fixed. Comparing complete OPD recipes answers a different question from disabling GOJO's two penalties under a common recipe; Appendix F describes the available comparisons.

E.6 Covariance Approximations

Local geometry and the old-student anchor. The minimum-credit-cost result applies to a conditional mean at a fixed query under a differentiable local probability perturbation. Equating this mean with v_{old} requires self-consistency; sampling endpoints with the old model alone does not ensure it. Without it, the covariance penalty centered at v_{old} is a surrogate. Independently chosen local credits also need not arise from one common endpoint reward. The local construction therefore does not measure KL between complete models or guarantee retention after a finite update.

Covariance Estimation. The implemented bank pools across queries, prompts, spatial positions, and a time interval. Its centered covariance includes variation in conditional-mean error as well as average local variability, as shown in Eq. (49). Channel, diagonal, and isotropic models then impose additional structure. Increasing the number of spatial observations does not remove these modeling assumptions or make those observations independent. Initializing an empty bank from the current batch adds dependence on that batch. Using previous statistics and resetting banks at each old-student refresh reduces, but does not eliminate, estimation error.

Interpretation of the experiments. Guidance, query sampling, RMS weighting, and unconditional consistency all influence the learned model in addition to the inner target. Their settings must be considered when interpreting complete-recipe comparisons. The reported covariance and protection studies use the measured settings described in Appendix F; they do not establish optimal hyperparameters, an advantage of isotropic covariance, or equivalence to a scalar update of the same norm.

F Experimental Protocols and Additional Analysis

We detail the experimental setting for two experiments. Specifically the Qwen-Image motivation example and the SD3.5-Medium experiments.

Table 3: Training recipes used in the SD3.5-Medium behavior diagnostics. CFG-U4 uses teacher as conditional branch and base model as unconditional branch, with CFG equals 4; self-CFG4 uses each teacher's own empty-condition field.

Recipe	Teacher field	Training queries	Conditional loss
DiffusionOPD	CFG-U4	Rollout states, uniform subset	MSE
DanceOPD	self-CFG4	Rollout states, Beta$(5, 2)$	MSE
GOJO	CFG-U4	$8\times$ Endpoint re-noising	MSE

F.1 Training setup and metric conventions

SD3.5-Medium training. The student and three teachers share the SD3.5-Medium base model. The teachers are optimized over rewards on GenEval, OCR, and PickScore. Each student variant (diffusionOPD, DanceOPD and GOJO) trains on all three source-task teachers. We tested both hard routing and soft-routing, where hard routing matches one teacher per epoch, and soft-routing averages all teachers at the same epoch.

Experimental setup. The diagnostic runs use 512×512 resolution images. We initialize the student model with the parameters of base model. We utilize rank-32 LoRA with scaling parameter 64, learning rate 10^{-5}, and no weight decay. We unify the OPD steps to be 1000. Training uses 8 workers. Each epoch performs one optimizer update and refreshes the frozen rollout student. We list the settting for four OPD methods in Table 3. On top of that, we unify the sampler with deterministic 40-step DPM2, and fix the student CFG to 1. GOJO constructs eight queries by re-noising the sampled endpoint on a separate time grid.

Evaluation Setups. Table 1 reports the main image evaluations, including the additional CLIPScore, HPSv2.1, aesthetic, and ImageReward measures. The component study instead uses training-time monitoring: 128 fixed prompts per task, evaluated every 20 updates. We report the mean reward over eleven checkpoints at optimization step $800, 820, \ldots, 1000$. In this monitoring protocol, GenEval is the continuous `eval/geneval/geneval` reward; separately logged accuracy and strict accuracy are distinct metrics. Repeated checkpoints from one training seed are not independent experimental replicates. Training diagnostics use updates 800–999 unless otherwise stated.

F.2 Qwen-Image motivation Example

Queries and gradient scope. Figure 2 uses trained checkpoints 100, 200, and 300, with texture and text specialists. Each checkpoint uses eight fixed prompts per task and the same seed. The image resolution is 1024×1024 and is sampled through deterministic 40-step DPM2 with CFG1. Both teachers are evaluated at the same intermediate rollout latent, irrespective of its prompt type or rollout owner. Student-, texture-teacher-, and text-teacher-owned rollouts yield 384 paired observations per checkpoint at steps $0, 11, 28, 35, 36, 37, 38, 39$. The main figure uses student-owned states, giving 64 queries per task and checkpoint.

Velocity, direct-gradient, and supervision-loss records are paired by task, prompt ID, seed, rollout owner, and denoising step, with matching prompt hashes and noise levels. Each conflict rate is the OPD-weighted fraction of negative per-query cosines, using the same queries and weights in the velocity and gradient panels. The sampled steps cover 34.17% of the full OPD time weight. Gradients use all coordinates in transformer blocks 0/8/16/23/31 and the output head. We measure per-query cosines on this parameter subset.

Teacher fidelity and numerical limits. Panel (c) averages the source-teacher OPD loss over the same queries and normalizes each task to checkpoint 100. The texture loss rises by 59.2%, while the text loss falls by 7.5%; these values measure uneven teacher matching and do not directly measure image quality or retention. The plotted rates and losses can be reconstructed from the records. We retain direct-gradient measurements but omit reconstructed component ratios.