FLOW-GUIDED WASSERSTEIN COMPENSATION DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-step flow distillation constructs training directions from model predictions at student-generated states, but these directions need not realize a prescribed evolution toward a reference distribution. We introduce Flow-Guided Wasserstein Compensation Distillation (FWCD), a framework that combines existing flow guidance with distributional information from unpaired reference samples. At each noise level, FWCD defines the minimum-energy correction to a scaled flow prior that completes a prescribed marginal response. A weak primal–dual formulation learns a scalar potential from unpaired samples; its input gradient supplies the correction for magnitude-controlled student updates. Under suitable function-space assumptions, we characterize the optimal field and its residual decomposition, relate population dual suboptimality to field approximation error, and establish Wasserstein contraction for the corresponding ideal distribution dynamics. Experiments on ImageNet-256 and SANA-1.6B demonstrate improvements over Mean Flow Distillation with four-step inference at matched student and auxiliary update counts. The flow-prior variant reduces ImageNet FID from 4.551 to 4.222 and SANA teacher-reference FID from 27.048 to 24.755. Ablations show task-dependent objective preferences: additive compensation achieves lower mean FID and KID on ImageNet at the default fitting budget, while flow-guided compensation performs better on SANA under a shared cap. FWCD preserves the student architecture and requires no additional model evaluations at inference.

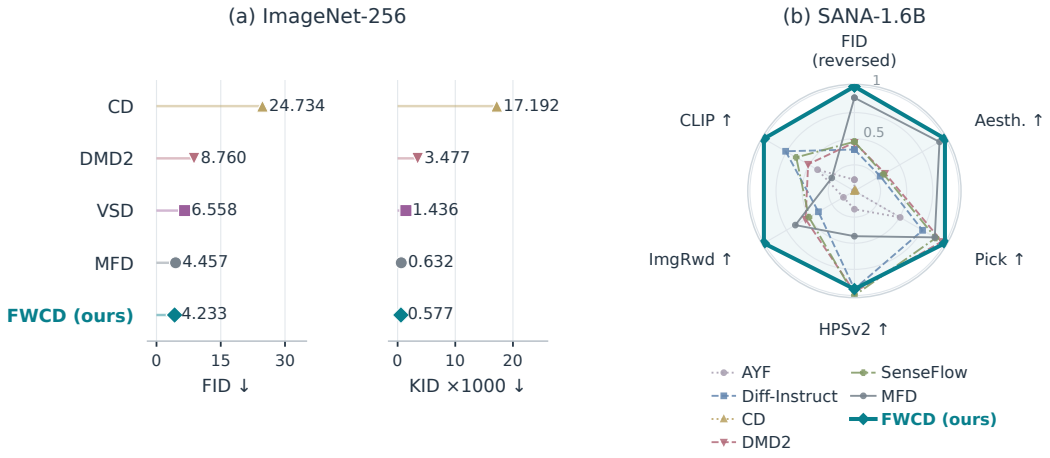

Figure 1: Four-step generation on (a) ImageNet-256 and (b) SANA-1.6B. SANA uses per-metric min–max scores over the seven methods, with FID reversed; outward is better.

1 INTRODUCTION

Few-step distillation compresses a multistep generator while preserving its output distribution (Lee et al., 2026; Zhao et al., 2026). Its training guidance must therefore combine the teacher's generative information with the discrepancy between the current student and target laws.

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.

A residual between teacher and student-auxiliary velocity predictions yields a base direction that can be evaluated at student query states (Zhao et al., 2026). That direction does not automatically satisfy a prescribed local density change between the current student marginal and a reference marginal. Both laws are typically available only through samples, so densities, divergences, and global optimal-transport couplings are poor objects for the training inner loop. A potential fitted from unpaired samples can estimate the response still missing from this base direction. This connects existing flow information with explicit marginal supervision.

We formulate FWCD as a constrained minimum-correction problem. Given a scaled flow prior, we seek the smallest correction that makes the total field realize a prescribed relaxation toward the reference marginal. The potential accounts for the response already supplied by the prior and learns the remainder through a sample-based weak dual. Correction energy is measured in the local Wasserstein kinetic geometry (Benamou & Brenier, 2000; Ambrosio et al., 2005).

Our contributions are the FWCD formulation of this constrained guidance problem; a pairing-free weak dual with a closed-form readout and a characterization of the ideal distributional dynamics; and controlled evidence on ImageNet-256 and SANA-1.6B at matched student-update budgets. Both tasks compare the flow-guided dual with an additive $\mathcal{D}_0$ control on $0.9b$. Objective, fitting budget, and magnitude rule yield different quality trade-offs.

2 Problem Setup

Student and target laws. The student generator produces $X^P = G_\theta(Z)$ with $Z \sim P_Z$ and $X^P \sim P_\theta$. The target law is written Q; in the main construction Q is the teacher generated law, and when real samples are used as the reference this is stated with the experiment. At noise level t, a student query is

$$X_t^P = a_t X^P + \sigma_t \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I),$$

with law p_t^θ. The reference query law is q_t. It may be formed from a teacher trajectory or by applying the same noise kernel to teacher or real endpoints; the two constructions are defined separately and are not assumed equal (Section 7.4).

Models and the base flow direction. The implementation uses a frozen teacher, the student generator G_θ, a student-velocity auxiliary model, and a scalar potential h_ω. Write $b_t(x)$ for the state-space direction of the existing distillation method. In the Mean Flow Distillation (MFD) instance used here,

$$b = u_{\text{aux}} - u_{\text{teacher}}. \tag{1}$$

Here u_{aux} and u_{teacher} are mean velocities from the current time t to an earlier time r. Writing $\Phi^j_{r\leftarrow t}(x, c)$ for the state of model j integrated from query x to time r,

$$u_j(x, t, r, c) = \begin{cases} \dfrac{x - \Phi^j_{r\leftarrow t}(x, c)}{t - r}, & r < t, \\ v_j(x, t, c), & r = t, \end{cases} \qquad j \in \{\text{aux}, \text{teacher}\}. \tag{2}$$

Experiments estimate this mean velocity by Euler integration on the same student query. The derivation treats b_t as an abstract prior; (1) is the realized MFD direction. The original method uses $g_t = b_t$. FWCD replaces g_t by a compensated field constructed from b_t and (p_t^θ, q_t).

Two times. The noise or generation time t is distinct from the ideal distributional evolution time τ. At fixed t, write $p_{\tau,t}$ for the evolving student query law and q_t for the fixed reference. A snapshot of these laws is written μ and ν, matching the appendix. The time τ is not an optimizer iteration.

3 Flow-Guided Wasserstein Compensation Distillation

Figure 2 summarizes the FWCD training and inference procedure.

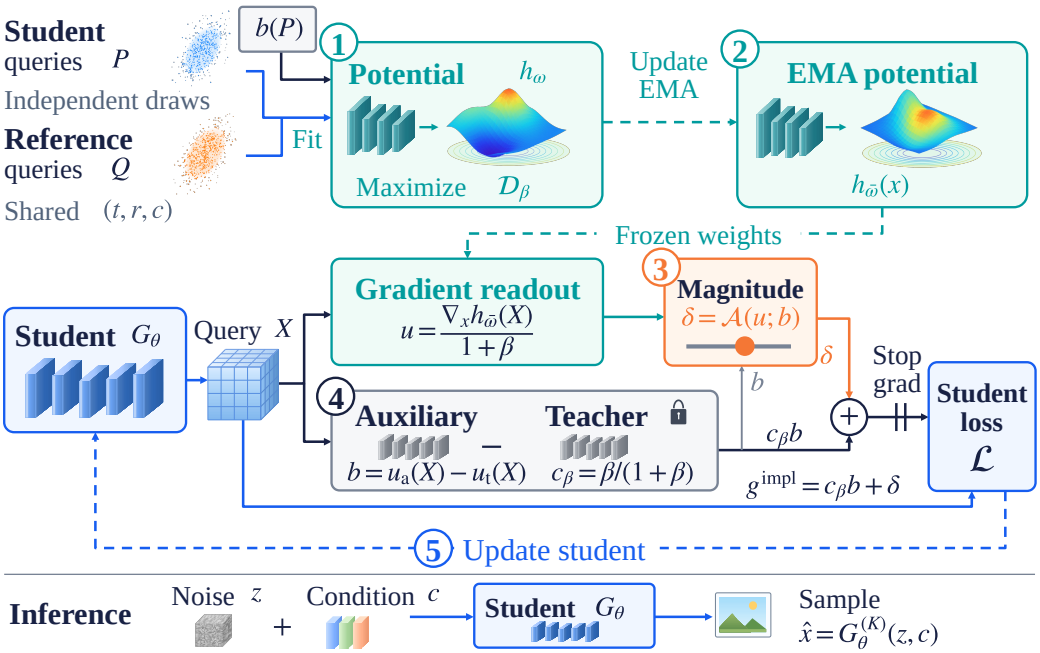

Figure 2: Overview of FWCD. **(1)** Fit the potential on independent queries with Eq. (10). **(2)** Update the potential EMA (Algorithm 1). **(3)** Read the correction from the EMA potential (Eq. (6)) and apply magnitude control (Eq. (12) for capped variants). **(4)** Form the base direction from auxiliary and teacher mean velocities (Eqs. (1)–(2)). **(5)** Combine the directions (Eq. (11)) and backpropagate through the student query with fixed guidance (Eq. (13)). Inference uses the student alone.

3.1 Prescribed density change

To specify what the compensated field should accomplish, we prescribe a linear relaxation of the student marginal toward the reference and express it as a constraint on the total guidance field. Fix the noise time t and let student particles evolve in τ along a guidance field,

$$\frac{\mathrm{d}X_{\tau,t}}{\mathrm{d}\tau} = g_{\tau,t}(X_{\tau,t}).$$

The continuity equation is $\partial_\tau p_{\tau,t} + \nabla \cdot (p_{\tau,t} g_{\tau,t}) = 0$. We take the linear relaxation $\partial_\tau p_{\tau,t} = \lambda_t(q_t - p_{\tau,t})$ with $\lambda_t > 0$, so the total field must satisfy

$$-\nabla \cdot (\mu g) = \lambda(\nu - \mu). \tag{3}$$

3.2 Minimum-energy compensation with a flow prior

The prescribed density change in (3) does not uniquely determine a field. We select the field closest to a scaled flow prior, yielding a minimum-energy correction. Instantaneous kinetic energy $\frac{1}{2}\mathbb{E}_\mu\|g\|^2$ measures the motion needed to realize the prescribed change. A quadratic flow prior $\frac{\beta}{2}\mathbb{E}_\mu\|g - b\|^2$ penalizes deviation from the existing distillation direction. The guidance field is

$$\begin{aligned} g^\star_\beta = \arg\min_g \quad & \frac{1}{2}\mathbb{E}_\mu\|g\|^2 + \frac{\beta}{2}\mathbb{E}_\mu\|g - b\|^2 \\ \text{s.t.} \quad & -\nabla \cdot (\mu g) = \lambda(\nu - \mu). \end{aligned} \tag{4}$$

Let $c_\beta = \beta/(1+\beta)$. Completing the square shows that (4) minimizes $\mathbb{E}_\mu\|g - c_\beta b\|^2$ over fields satisfying the prescribed density response. Thus, with $u = g - c_\beta b$, FWCD selects the minimum-energy correction to the scaled flow prior. The potential accounts for the response already supplied by that prior and completes the remainder. Appendix C gives the equivalent correction problem and proof.

3.3 WEAK FORM

We access the query distributions through samples, without evaluating densities or divergences. We rewrite the density constraint as a weak relation between expectations, without requiring sample-to-sample matching. For admissible test functions f, integration by parts with the boundary conditions of Appendix A turns (3) into

$$\mathbb{E}_\mu\langle g, \nabla f\rangle = \lambda\big(\mathbb{E}_\nu f - \mathbb{E}_\mu f\big). \tag{5}$$

The scalar potential h is the dual variable of this weak constraint.

3.4 DUAL POTENTIAL

To solve the constrained field problem, we derive a scalar-potential dual and recover the total guidance field from the potential's input gradient and the base flow. We omit subscripts at fixed t, τ and introduce a scalar multiplier for the weak constraint (full Lagrangian in Appendix C). Stationarity in g at fixed h gives $g + \beta(g - b) - \nabla h = 0$, hence

$$g_\beta[h] = \frac{\beta b + \nabla h}{1 + \beta}. \tag{6}$$

Substituting and dropping h-independent constants yields the weak dual

$$\mathcal{D}_\beta(h) = \lambda\big(\mathbb{E}_\nu[h] - \mathbb{E}_\mu[h]\big) - \frac{1}{2(1+\beta)}\mathbb{E}_\mu\|\nabla h + \beta b\|^2. \tag{7}$$

The potential is trained by maximizing (7). The reference marginal enters through the contrast of expectations, and the flow prior through the gradient-energy term. At $\beta = 0$, this is the basic dual

$$\mathcal{D}_0(h) = \lambda\big(\mathbb{E}_\nu[h] - \mathbb{E}_\mu[h]\big) - \frac{1}{2}\mathbb{E}_\mu\|\nabla h\|^2. \tag{8}$$

3.5 CLOSED-FORM COMPENSATION

To connect the field readout to the implemented update, we distinguish the correction relative to the full base from the scaled-prior correction used for magnitude control. We write $g = b + r_{\text{comp}}$; combining this with (6) gives

$$r_{\text{comp}}(h) = \frac{\nabla h - b}{1 + \beta}. \tag{9}$$

The minimum-energy correction is defined relative to the scaled prior $c_\beta b$, with readout $u_\beta[h] = \nabla h/(1+\beta)$. The correction relative to the full base satisfies $r_{\text{comp}} = u_\beta[h] - b/(1+\beta)$. Section 4 applies magnitude control to $u_\beta[h]$.

4 LEARNING AND STUDENT UPDATE

Potential learning. To learn the correction from finite query batches, we estimate the dual objective with sample averages and add a gauge penalty on the potential's additive offset. Independent samples $x_i^P \sim \mu$ and $x_j^Q \sim \nu$ estimate

$$\widehat{\mathcal{D}}_\beta(h_\omega) = \lambda\Big(\frac{1}{N_Q}\sum_j h_\omega(x_j^Q) - \frac{1}{N_P}\sum_i h_\omega(x_i^P)\Big) - \frac{1}{2(1+\beta)N_P}\sum_i \|\nabla_x h_\omega(x_i^P) + \beta b(x_i^P)\|^2,$$

with $b_P = b(P)$ from Eq. (1) and the same time and condition context on both sides. The potential loss is

$$\mathcal{L}_h(P, Q, b_P) = s_{\text{loss}}\left[-\widehat{\mathcal{D}}_\beta(h_\omega; P, Q, \text{sg}(b_P)) + 10^{-4}\left(\frac{\overline{h}_P + \overline{h}_Q}{2}\right)^2\right]. \tag{10}$$

Student queries, reference queries, and b_P are stopped-gradient with respect to the generative models; the energy term is differentiated through $\nabla_x h_\omega$. Architecture, s_{loss}, warmup, M, and EMA settings are in Appendix E.

Algorithm 1 Flow-Guided Wasserstein Compensation Distillation

Require: Frozen teacher T; student G_θ; auxiliary A_ψ; potential h_ω; reference sampler; schedule; magnitude rule $\mathcal{A}$

Set $\beta = 9, \lambda = 1, W = 20, M = 5, \alpha = 2^{-1/8}$. Warm up ω for W steps on detached (P, Q, b_P) with (10); $\bar{\omega} \leftarrow \omega$.
for global step $n = 1, \ldots, N$ **do**
  if auxiliary update **then**
    Update ψ with the MFD auxiliary loss on detached student samples.
  else
    Sample (t, r, c); form differentiable query X^θ and detached fitting pair (P, b_P).
    for $j = 1, \ldots, M$ **do**
      Sample or reuse independent Q; update ω with $\mathcal{L}_h(P, Q, \mathrm{sg}(b_P))$.
    end for
    $\bar{\omega} \leftarrow \alpha\bar{\omega} + (1-\alpha)\omega$, $b \leftarrow \mathrm{sg}(u_{\mathrm{aux}}(X^\theta) - u_{\mathrm{teacher}}(X^\theta))$, $u \leftarrow \mathrm{sg}(\nabla_x h_{\bar{\omega}}(\mathrm{detach}(X^\theta))/10)$.
    $\delta \leftarrow \mathcal{A}(u; b)$ (permute within context for FWCD-Shuffle); $g^{\mathrm{impl}} \leftarrow 0.9b + \delta$; update θ with (13) using the queries retained for backpropagation.
  end if
end for
return Four-step student G_θ.

Magnitude control. To control the correction supplied during training, we apply a magnitude rule to the EMA potential readout and combine the result with the scaled base. With $h = h_{\bar{\omega}}$, the implemented field is

$$g^{\mathrm{impl}} = c_\beta b + \mathcal{A}(u_\beta[h]; b). \tag{11}$$

Natural magnitude is $\mathcal{A}(u; b) = u$. A relative RMS cap takes the form

$$\mathcal{A}(u; b) = \min\left\{1, \frac{\kappa\,\mathrm{RMS}(b)}{\mathrm{RMS}(u)}\right\} u, \tag{12}$$

where $\kappa > 0$ is the relative RMS limit; zero correction remains zero. At $\beta = 9$, the direction is $0.9b + \mathcal{A}(\nabla h/10; b)$. Active capping changes the prescribed density response even for an exact potential (Appendix C.2).

Student update. To translate state-space guidance into parameter updates, we use a stopped-gradient linear surrogate that backpropagates through the student query while holding the guidance fixed:

$$\mathcal{L}_{\mathrm{student}}(\theta) = -\frac{1}{B_{\mathrm{s}} d}\sum_{i=1}^{B_{\mathrm{s}}} \langle X^\theta_{t,i}, \mathrm{sg}(g^{\mathrm{impl}}_{t,i})\rangle, \qquad \nabla_\theta \mathcal{L}_{\mathrm{student}} = -\frac{1}{B_{\mathrm{s}} d}\sum_{i=1}^{B_{\mathrm{s}}} J_\theta(X_{t,i})^\top \mathrm{sg}(g^{\mathrm{impl}}_{t,i}), \tag{13}$$

where d is the query dimension and B_{s} counts queries retained for student backpropagation. Thus the mean covers both queries and latent coordinates. On SANA, $B_{\mathrm{s}} = 1$, while potential fitting uses 16 queries. Algorithm 1 specifies the warmup, alternating updates, and EMA schedule for the default $\beta = 9$ implementation.

5 Theoretical properties

To establish the theoretical properties of the compensation, we characterize the optimal population field under the stated assumptions and its corresponding ideal distributional dynamics. Proofs are in Appendices A–C.

Theorem 1 (Primal–dual characterization)**.** *Under a weighted Poincaré inequality, $\nu - \mu \in \dot{H}^{-1}(\mu)$, and the boundary or decay conditions of Appendix A, problem* (4) *at $\beta = 0$ has a unique minimum-energy solution $g_0^\star = \nabla h_0^\star$. The potential $h_0^\star$, unique modulo constants, satisfies* (5) *for all $f \in \dot{H}^1(\mu)$ and maximizes $\mathcal{D}_0$.*

Proposition 2 (Optimal field with a flow prior)**.** *Under Theorem 1, let $\beta \geq 0$ and $b \in L^2(\mu; \mathbb{R}^d)$. Write $b = P^\mu_\nabla b + b_\perp$, where P^μ_∇ is the $L^2(\mu)$-orthogonal projection onto the closure of admissible*

gradient fields. The solution of (4) *is*

$$g_\beta^\star = g_0^\star + \frac{\beta}{1+\beta} b_\perp = \frac{\beta b + \nabla h_\beta^\star}{1+\beta}, \tag{14}$$

and $-\nabla \cdot (\mu g_\beta^\star) = \lambda(\nu - \mu)$. *Equivalently,*

$$\frac{\nabla h_\beta^\star}{1+\beta} = g_0^\star - c_\beta P_\nabla^\mu b, \qquad c_\beta = \frac{\beta}{1+\beta}. \tag{15}$$

Every ideal total field satisfies $P_\nabla^\mu g_\beta^\star = g_0^\star$.

Equation (15) identifies the minimum-energy correction after accounting for the scaled prior's density response. Under exact population fitting in a fixed closed linear gradient class, residual fitting avoids duplicating representable base responses. The base can also supply components outside that class; whether this helps depends on their direction and magnitude relative to the missing target response (Appendix C.3). Every exact total field realizes the same prescribed relaxation, satisfying $W_2^2(p_{\tau,t}, \nu) \le e^{-\lambda_t \tau} W_2^2(\mu, \nu)$ (Proposition 3). These guarantees concern population fields before magnitude processing; finite fitting and parameterized student updates are evaluated experimentally.

6 Related work

Flow distribution distillation and reference supervision. Mean Flow Distillation and Decoupled MeanFlow enable few-step sampling from pretrained flow models (Zhao et al., 2026; Lee et al., 2026). ScoreLiDAR relates student-induced query mismatch to unreliable guidance (Zhang et al., 2025). DMD2 combines two-timescale optimization with real-data GAN supervision (Yin et al., 2024). LADD uses teacher-generated latents and teacher features for adversarial distillation, while SDXL-Lightning trains LoRA students in progressive adversarial stages (Sauer et al., 2024; Lin et al., 2024). Adversarial objectives also supply input-gradient directions. The distinction of FWCD is its prescribed marginal response and prior-guided variational field, rather than the use of a direction instead of a loss. Consistency Models and Latent Consistency Models learn trajectory-consistent mappings (Song et al., 2023; Luo et al., 2023a); InstaFlow straightens trajectories before distillation (Liu et al., 2024).

Sobolev distributional transport. Sobolev GAN defines a discrepancy through a gradient-energy constraint (Mroueh et al., 2018). Sobolev Descent uses the critic gradient to transport a distribution, with weighted Poisson and minimum-energy interpretations (Mroueh et al., 2019). With compatible spaces and boundary conditions, $\mathcal{D}_0$ is the quadratic-penalty form of this discrepancy. Action Matching learns a scalar potential whose gradient transports one sample population to another over time (Neklyudov et al., 2023); it targets the dynamics between observed marginals, whereas the FWCD potential is a compensation at a fixed noise time relative to a given flow direction. Including b in the dual changes the potential's fitting target; $\beta = 0$ recovers the basic dual.

Coupling geometry. Flow Matching, Rectified Flow, and Stochastic Interpolants use endpoint couplings to construct paths and regression targets (Lipman et al., 2023; Liu et al., 2023; Albergo & Vanden-Eijnden, 2023). Data-dependent and minibatch-OT couplings modify this geometry (Albergo et al., 2024; Tong et al., 2024). The kinetic formulation specifies density change through a tangent field (Ambrosio et al., 2005; Villani, 2009). Flow-matching approximation errors are studied by Benton et al. (2024); Zhou & Liu (2025). Appendix J compares the kinetic construction with minibatch pairing displacements.

7 Experiments

7.1 Experimental setup

In the MFD-based fixed-update comparisons, methods share student and auxiliary update budgets and use four-step inference. ImageNet-256 uses a pretrained multistep SiT-XL/2 teacher (Ma et al., 2024) and a four-step LoRA student with rank/alpha $64/64$. A global batch contains 32 queries from two classes. Training alternates 50 auxiliary and 20 student updates for 2000 global steps. Default-protocol tables average R1–R3; replay comparisons average R4 and R5 (Table 5). SANA-1.6B (Xie

et al., 2025) uses 1024-pixel generation with separate rank/alpha $64/64$ LoRAs; 3000 global steps give 2728 auxiliary and 272 student updates. ImageNet uses the relative cap with $\kappa = 0.1$; SANA uses natural magnitude by default and $\kappa = 0.1$ for capped variants. Evaluation and query construction are specified below and in Appendix E.

7.2 Few-step generation

ImageNet-256. Potential fitting uses independently noised student endpoints and real-image VAE latents,

$$x_t^P = (1-t)G_\theta^{(4)}(\epsilon, c) + t\xi_P, \qquad x_t^Q = (1-t)z_{\text{real}}^c + t\xi_Q, \tag{16}$$

with independent noise; potential-fitting and student-update queries are generated separately. Table 1 compares MFD, Attenuation, the flow-guided dual FWCD (β=9), its shuffle, and the additive $\mathcal{D}_0$ control. For MFD and its variants, checkpoint 2000 generates 50k images with four Euler steps, BF16, and CFG 1. FID uses Inception features, and KID is reported $\times 1000$. A same-seed comparison with DMD2, VSD, and CD appears in Table 6.

Table 1: ImageNet-256 four-step generation. Mean $\pm$ sample standard deviation over R1–R3, with 560 student updates and 50k evaluation images per run. Potential arms use $0.9b$, a 10% cap, and $M = 5$.

Method	FID ↓	KID $\times 1000$ ↓
MFD	4.551 ± 0.170	0.663 ± 0.054
Attenuation	4.549 ± 0.237	0.735 ± 0.095
Additive $\mathcal{D}_0$	$\mathbf{4.059} \pm 0.163$	$\mathbf{0.541} \pm 0.015$
FWCD (β=9)	4.222 ± 0.105	0.592 ± 0.019
FWCD-Shuffle	4.442 ± 0.372	0.616 ± 0.094

Flow-guided FWCD (β=9) lowers mean FID and KID over MFD by 7.22% and 10.79%; Additive $\mathcal{D}_0$ on the same mixture and cap reaches 10.81% and 18.40%. Both potential arms lie below Attenuation and FWCD-Shuffle on these two means. R3 favors $\mathcal{D}_\beta(\beta$=9) in FID and additive $\mathcal{D}_0$ in KID. Attenuation leaves mean FID and KID unchanged or worse, so attenuation alone does not account for these gains. Shuffling the corrections recovers part of the gain. Figure 3 shows samples spaced evenly in the full evaluation list. Early-only compensation for the first 140 student updates, followed by MFD, remains below MFD on the R1–R3 means (FID 4.313, KID 0.610). Full FWCD (β=9) has lower mean FID and KID (Table 8). The early window is not a general stopping rule: R3 is worse on both metrics than full compensation.

SANA-1.6B. Student endpoints are independently re-noised, while reference queries come from teacher trajectories:

$$x_t^P = (1-t)G_\theta^{(4)}(\epsilon, c) + t\xi_P, \qquad x_t^Q = \Phi_{t\leftarrow 1}^{\text{teacher}}(\xi_Q, c). \tag{17}$$

The teacher trajectory uses CFG 4.5 and Euler step bound 0.03. The potential uses 16 queries per prompt. Table 2 reports four-step SANA generation scores. Appendix H gives our fixed-5k four-step evaluation. Our students use checkpoint 3000, NFE 4 and CFG 1, with NFE-20/CFG-4.5 teacher reference images.

Within our four-step evaluation, FWCD improves all six scores in Table 2 over MFD: FID falls by 8.48%, and HPSv2 and ImageReward rise by 3.84% and 19.54%. Natural-magnitude shuffling loses most of the FID gain (28.250). Capping improves FID and four preference scores; natural magnitude retains the highest CLIPScore among the potential variants. Capped shuffling raises FID by 0.40. Additive $\mathcal{D}_0$ under the cap is worse than FWCD (cap) on all six metrics. Appendix H reports the correction-magnitude diagnostics.

COCO transfer. We compare the existing MFD and capped FWCD SANA checkpoints on COCO2014 with shared prompts and noise (Appendix K). On 30k images, capped FWCD lowers real-reference FID from 27.470 to 20.729 and raises CLIPScore from 30.497 to 30.866. On a shared 5k subset, FWCD improves all five preference scores over MFD.



Figure 3: Four-step ImageNet-256 samples with matched class and initial noise across rows. Shuffle denotes FWCD-Shuffle.

Table 2: SANA-1.6B four-step generation.

Method	NFE	FID$^{\downarrow}$	Aesthetic$^{\uparrow}$	PickScore$^{\uparrow}$	HPSv2$^{\uparrow}$	ImageReward$^{\uparrow}$	CLIPScore$^{\uparrow}$
AYF (2025)	4	43.68	6.231	0.1963	0.2298	0.302	29.176
Diff-Instruct (2023b)	4	37.55	6.453	0.1985	0.2434	0.368	29.393
CD (2023)	4	45.97	6.242	0.1918	0.2267	0.273	28.925
DMD2 (2024)	4	36.13	6.495	0.2004	0.2439	0.402	29.240
SenseFlow (2026)	4	35.97	6.482	0.1998	0.2445	0.393	29.321
MFD	4	27.05	6.964	0.1997	0.2344	0.429	29.079
Attenuation	4	28.75	6.973	0.1997	0.2338	0.456	29.130
Additive $\mathcal{D}_0$ (cap)	4	26.68	7.028	0.1999	0.2409	0.476	29.261
Additive $\mathcal{D}_0$-Shuffle (cap)	4	28.56	7.099	0.1994	0.2437	0.477	29.094
FWCD	4	24.76	7.011	0.2007	0.2434	0.512	**29.541**
FWCD-Shuffle	4	28.25	7.049	0.2000	0.2374	0.439	29.356
FWCD (cap)	4	**23.47**	**7.145**	**0.2010**	**0.2465**	**0.579**	29.324
FWCD-Shuffle (cap)	4	23.88	6.965	0.2005	0.2413	0.525	29.270

7.3 Guidance and compensation design

The controls separate the potential objective, correction assignment, and base scaling. Additive $\mathcal{D}_0$ fits (8) but supplies $0.9b + \mathcal{A}(\nabla h; b)$, retaining the same base as $\mathcal{D}_\beta(\beta{=}9)$. FWCD-Shuffle reassigns that arm's corrections within each class or prompt. Attenuation uses $0.9b$ alone. These comparisons isolate changes to the implemented correction on a shared scaled base.

A separate replay comparison fixes the correction-to-base RMS sequence across potential arms. Across R4 and R5, removing the marginal contrast from $\mathcal{D}_\beta(\beta{=}9)$ worsens both metrics (Table 10). This establishes that term's contribution in the tested residual configuration; the changed target dynamics do not isolate network capacity.

Under the common cap protocol, the Shapley decomposition assigns positive FID contributions to preserving correspondence in both tasks, but opposite signs to switching from $\mathcal{D}_0$ to $\mathcal{D}_\beta(\beta{=}9)$ (Appendix I.4).

7.4 FITTING BUDGET AND REFERENCE CONSTRUCTION

The number of potential updates M before each student update is the potential's fitting budget on the current (μ_t, ν_t, b) (Appendix I.3). On ImageNet, $M = 1$ lowers the mean FID and KID of the residual objective relative to $M = 5$, but R3 is worse on both metrics. On SANA, one update retains the FID gain over MFD while all five preference scores are lower than at $M = 5$; $M = 10$ lowers FID further, while HPSv2 peaks at $M = 5$ (Table 14). Table 3 compares both objectives: additive $\mathcal{D}_0$ has lower mean FID and KID at $M = 5$, whereas $\mathcal{D}_\beta(\beta{=}9)$ is lower at $M = 1$. The KID ordering reverses in all three paired runs. These ablations favor different budgets across objectives and metrics; the main comparison keeps the common $M = 5$ setting.

Table 3: ImageNet objective and fitting budget (10% cap). Mean ± sample standard deviation over R1–R3, 50k evaluation.

Objective	M	FID $^{\downarrow}$	KID $\times 1000^{\downarrow}$
Additive $\mathcal{D}_0$	5	$\mathbf{4.059} \pm 0.163$	$\mathbf{0.541} \pm 0.015$
FWCD (β=9)	5	4.222 ± 0.105	0.592 ± 0.019
Additive $\mathcal{D}_0$	1	4.134 ± 0.282	0.732 ± 0.071
FWCD (β=9)	1	4.061 ± 0.065	0.560 ± 0.072

The asymmetric SANA construction in (17) can retain a query-law gap even when the student matches the teacher endpoint (Appendix D.2). A 64-prompt teacher self-comparison, with 16 samples per prompt, yields MMD intervals spanning zero and does not establish equivalence. In a separate reference-construction comparison, 3000-step endpoint re-noising gives FID 25.359, versus 27.930 for its paired trajectory-reference run; four of six scores improve (Appendix D.3).

8 LIMITATIONS

Our guarantees concern population fields at fixed query laws; finite potential fitting, magnitude control, and parameterized student updates need not realize the ideal relaxation. FWCD adds potential optimization and reference construction, so matched model-update counts do not establish an equal-compute advantage. SANA results, including COCO transfer, use one training run per configuration and do not quantify training variability. Experiments cover four-step LoRA students, with task- and metric-dependent objective and budget trade-offs. Reference construction determines the compensation target, and marginal supervision does not directly enforce local image structure. Qualitative failures are reported in Appendix J.4.

9 CONCLUSION

FWCD connects a prescribed marginal response to a correction that can be fitted from independent samples. In the evaluated default configurations, the correction improves on MFD, while the preferred objective and fitting budget depend on the task and metric. ImageNet favors Additive $\mathcal{D}_0$ in the default $M = 5$ means; SANA benefits from the flow-guided objective. These results support treating the potential objective, correction assignment, and magnitude rule as separate design choices. The learned potential is needed only during training.

ETHICS AND SOCIETAL IMPACT

FWCD modifies training guidance for existing image generators and retains the student architecture at inference. Potential benefits concern generation quality without an additional inference model; overall training costs increase through potential fitting and reference construction. The method does not directly address biases, privacy concerns, or harmful and deceptive image generation inherited from the teacher and data. Our experiments evaluate image quality and automated preference scores, not safety or fairness. Any release should respect the model and data licenses and be accompanied by appropriate usage restrictions and content safeguards.

AI USE STATEMENT

In this work, we used generative AI tools to assist with experiment code, analysis, plotting, and manuscript drafting. We did not use them to generate data or the reported experimental results. We reviewed the AI-assisted code and text, and we take responsibility for the contents of this paper.

REPRODUCIBILITY STATEMENT

The field assumptions and proofs are in Appendices A–D. Algorithm 1 and Appendix E specify query construction, optimization, sample reuse, and evaluation. Appendix F records replicate IDs and random settings. Appendices G–K report quality results, Shapley decompositions, pairing controls, and the COCO protocol.

REFERENCES

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.15571.

Michael S. Albergo, Mark Goldstein, Nicholas M. Boffi, Rajesh Ranganath, and Eric Vanden-Eijnden. Stochastic interpolants with data-dependent couplings. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 921–937, 2024. URL https://proceedings.mlr.press/v235/albergo24a.html.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser, 2005.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=uqQPyWFDhY.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1801.01401.

Xingtong Ge, Xin Zhang, Tongda Xu, Yi Zhang, Xinjie Zhang, Yan Wang, and Jun Zhang. SenseFlow: Scaling distribution matching for flow-based text-to-image distillation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2506.00523.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.08500.

Kyungmin Lee, Sihyun Yu, and Jinwoo Shin. Decoupled MeanFlow: Turning flow models into flow maps for accelerated sampling. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/699124f548e4cb86674b20eae0fd8619-Abstract-Conference.html.

Shanchuan Lin, Anran Wang, and Xiao Yang. SDXL-Lightning: Progressive adversarial diffusion distillation, 2024. URL https://arxiv.org/abs/2402.13929.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

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.

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.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. InstaFlow: One step is enough for high-quality diffusion-based text-to-image generation. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.06380.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023a. URL https://arxiv.org/abs/2310.04378.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-Instruct: A universal approach for transferring knowledge from pre-trained diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f115f619b62833aadc5acb058975b0e6-Abstract-Conference.html.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2401.08740.

Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev GAN. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1711.04894.

Youssef Mroueh, Tom Sercu, and Anant Raj. Sobolev descent. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2976–2985, 2019. URL https://proceedings.mlr.press/v89/mroueh19a.html.

Kirill Neklyudov, Rob Brekelmans, Daniel Severo, and Alireza Makhzani. Action matching: Learning stochastic dynamics from samples. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25858–25889, 2023. URL https://proceedings.mlr.press/v202/neklyudov23a.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your flow: Scaling continuous-time flow map distillation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d79ac139911df25d27f14bbf008deaee-Abstract-Conference.html.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11. Association for Computing Machinery, 2024. doi: 10.1145/3680528.3687625. URL https://doi.org/10.1145/3680528.3687625.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023. URL https://arxiv.org/abs/2303.01469.

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. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2302.00482.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. ProlificDreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1a87980b9853e84dfb295855b425c262-Abstract.html.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution image synthesis with linear diffusion transformers. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.10629.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.14867.

Shengyuan Zhang, An Zhao, Ling Yang, Zejian Li, Chenye Meng, Haoran Xu, Tianrun Chen, AnYang Wei, Perry Pengyun Gu, and Lingyun Sun. Distilling diffusion models to efficient 3D LiDAR scene completion. In *International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2412.03515.

An Zhao, Shengyuan Zhang, Zhongjian Sun, Yixiang Zhou, Zejian Li, Ling Yang, Tianrun Chen, and Lingyun Sun. Mean flow distillation: Robust and stable distillation for flow matching models. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.11155.

Zhengyu Zhou and Weiwei Liu. An error analysis of flow matching for deep generative modeling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 78903–78932, 2025. URL https://proceedings.mlr.press/v267/zhou25l.html.

Appendices A–D give the weak form, flow-prior compensation, and reference laws for Flow-Guided Wasserstein Compensation Distillation (FWCD). Appendix E specifies implementation and evaluation. Appendix F records replicate IDs and random settings. Appendices G–I collect full quality results and potential ablations. Appendix J reports pairing controls and failure cases; Appendix K gives the COCO protocol and scores.

A Notation and weak formulation

The field problems are defined on $\Omega = \mathbb{R}^d$, a flat torus, or a bounded Lipschitz domain. Probability measures belong to $\mathcal{P}_2(\Omega)$. We use $L^2(\mu; \mathbb{R}^d)$ for square-integrable vector fields, G_μ for the closure of admissible smooth gradients, and $N_\mu = G_\mu^\perp$ for the weighted divergence-free subspace. The weighted Sobolev space $\dot{H}^1(\mu)$ identifies functions differing by a constant and has norm $\|\nabla h\|_{L^2(\mu)}$. Its continuous dual is $\dot{H}^{-1}(\mu)$. A weighted Poincaré inequality controls a mean-zero representative by its gradient.

The density constraint is interpreted weakly:

$$\int \langle \nabla f, g \rangle \, \mathrm{d}\mu = \lambda \langle \nu - \mu, f \rangle_{\dot{H}^{-1}, \dot{H}^1}, \qquad f \in \dot{H}^1(\mu). \tag{18}$$

For integrable representatives, the right side equals $\lambda(\mathbb{E}_\nu f - \mathbb{E}_\mu f)$. We use periodic boundaries on the torus, zero flux on bounded domains, and sufficient decay on $\mathbb{R}^d$ to remove boundary terms.

For observed conditions, define $\mu^c, \nu^c, h(x, t, r, c)$ and average the dual over c:

$$\mathbb{E}_c \left[\lambda(\mathbb{E}_{\nu^c} h - \mathbb{E}_{\mu^c} h) - \frac{1}{2(1+\beta)} \mathbb{E}_{\mu^c} \|\nabla h + \beta b\|^2 \right]. \tag{19}$$

A mixed-condition batch estimates this expectation over query laws within observed conditions.

Table 4: Principal notation. Noise time, ideal density-evolution time, and student-update index are separate quantities.

Symbol	Meaning
t, r, c	Noise-time pair and observed class or text condition
μ, ν	Student and reference query laws at fixed context
τ, k, n	Ideal evolution time, student-update index, global model-update index
b, δ	Base direction and implemented correction component
$g^\star_\beta, g_\beta[h], g^{\mathrm{impl}}$	Ideal constrained field, potential readout, implemented student direction
$\omega, \bar{\omega}$	Potential parameters and their EMA
θ, ψ	Student and auxiliary trainable parameters
β, λ	Flow-prior weight and density-relaxation rate
M	Potential updates before each student update
κ	Relative RMS limit for the capped correction
r_k	Correction-to-base RMS ratio in the relative-magnitude ablation
P^μ_∇	Orthogonal projection onto G_μ

Fixed-noise well-posedness. At a fixed context, λ denotes λ_t. Suppose $\nu \ll \mu$, $w = \mathrm{d}\nu/\mathrm{d}\mu - 1 \in L^2(\mu)$, and $\mathrm{Var}_\mu(f) \le C_P \|\nabla f\|^2_{L^2(\mu)}$. Since $\mathbb{E}_\mu w = 0$, Cauchy–Schwarz and Poincaré give

$$|\lambda(\mathbb{E}_\nu f - \mathbb{E}_\mu f)| \le \lambda\sqrt{C_P}\|w\|_{L^2(\mu)}\|\nabla f\|_{L^2(\mu)}, \qquad \|g^\star_0\|_{L^2(\mu)} \le \lambda\sqrt{C_P}\|w\|_{L^2(\mu)}.$$

These are sufficient conditions and an upper bound. Gaussian smoothing gives a regular positive density but does not by itself bound these constants uniformly over noise levels. Low-density bottlenecks can worsen conditioning; a worsening upper bound does not prove divergence of the field energy. This issue is distinct from the shared-kernel identifiability result in Appendix D.

B Minimum-energy correction

Define the bounded linear functional $\ell(f) = \lambda\langle \nu - \mu, f\rangle_{\dot{H}^{-1},\dot{H}^1}$. By Riesz representation on the weighted Sobolev Hilbert space, there is a unique $h^\star$ modulo constants, written $h^\star_0$ below, satisfying

$$\int \langle \nabla h^\star_0, \nabla f\rangle \,\mathrm{d}\mu = \ell(f) \quad \text{for all } f \in \dot{H}^1(\mu). \tag{20}$$

Therefore $g^\star_0 = \nabla h^\star_0$ satisfies the density constraint. Every other feasible field is $g^\star_0 + n$, $n \in N_\mu$. Orthogonality gives

$$\|g^\star_0 + n\|^2_{L^2(\mu)} = \|g^\star_0\|^2_{L^2(\mu)} + \|n\|^2_{L^2(\mu)},$$

which proves uniqueness and minimum energy. Completing the square yields

$$\ell(h) - \frac{1}{2}\|\nabla h\|^2_{L^2(\mu)} = \frac{1}{2}\|\nabla h^\star_0\|^2_{L^2(\mu)} - \frac{1}{2}\|\nabla h - \nabla h^\star_0\|^2_{L^2(\mu)}. \tag{21}$$

The maximum is attained at $h^\star_0$ and equals the primal minimum, proving Theorem 1.

For comparison with Sobolev distributional transport, define in this same space

$$S(\nu, \mu) = \sup_{\|\nabla f\|_{L^2(\mu)} \le 1} \langle \nu - \mu, f\rangle.$$

Optimizing separately over the amplitude of f gives

$$\sup_h \mathcal{D}_0(h) = \frac{\lambda^2}{2} S(\nu, \mu)^2. \tag{22}$$

This is the quadratic-penalty counterpart of the source-weighted Sobolev discrepancy (Mroueh et al., 2019). Equality requires the same linear function space, gradient weighting, and boundary conditions.

Let ρ satisfy the assumptions of Appendix A and let $a \in L^2(\rho; \mathbb{R}^d)$. For probability laws ν_ε satisfying the uniform bound

$$|\langle \nu_\varepsilon - \rho, f\rangle - \varepsilon\mathbb{E}_\rho\langle \nabla f, a\rangle| \le C\varepsilon^2\|\nabla f\|_{L^2(\rho)}, \qquad f \in \dot{H}^1(\rho),$$

continuity of the Riesz map gives

$$g_{0,\varepsilon} = \lambda\varepsilon P^{\rho}_{\nabla} a + O_{L^2(\rho)}(\varepsilon^2). \tag{23}$$

Thus the weak-dual field recovers the gradient component responsible for the prescribed first-order density change.

Proposition 3 (Ideal relaxation dynamics)**.** *Fix the noise time and let μ_τ solve $\partial_\tau \mu_\tau = \lambda(\nu - \mu_\tau)$ with $\mu_0 = \mu$, where μ, ν have finite second moments. Then*

$$\mu_\tau = e^{-\lambda\tau}\mu + (1 - e^{-\lambda\tau})\nu, \qquad W_2^2(\mu_\tau, \nu) \le e^{-\lambda\tau} W_2^2(\mu, \nu).$$

For a discrete step $\bar{\mu}_{k+1} = (1-\eta\lambda)\mu_k + \eta\lambda\nu$ with $0 \le \eta\lambda \le 1$, $W_2^2(\bar{\mu}_{k+1}, \nu) \le (1-\eta\lambda) W_2^2(\mu_k, \nu)$.

Proof. The equation is linear in μ_τ with constant source $\lambda\nu$, which gives the mixture form. For $s \in [0, 1]$ and $\mu_s = (1 - s)\mu + s\nu$, let π be an optimal coupling of (μ, ν) and let Δ be the diagonal coupling of (ν, ν). The mixture $(1 - s)\pi + s\Delta$ couples μ_s with ν, and its cost is $(1 - s)W_2^2(\mu, \nu)$. Taking the infimum over couplings gives $W_2^2(\mu_s, \nu) \le (1 - s)W_2^2(\mu, \nu)$. The continuous statement uses $s = 1 - e^{-\lambda\tau}$ and the discrete statement uses $s = \eta\lambda$. □

The proposition applies to any field satisfying the constraint, including $g_0^\star$ and the prior-compensated $g_\beta^\star$ of Appendix C, and identifies the constraint as a first-order contraction of the student law toward the reference law. A realized student replaces this ideal step by finite-sample potential fitting, the magnitude rule, the student Jacobian, and the optimizer; we do not bound that error.

C Flow prior and residual compensation

Let $c_\beta = \beta/(1 + \beta)$ and $\|w\|_\mu^2 = \mathbb{E}_\mu\|w\|^2$. Completing the square in the objective of (4) gives

$$\frac{1}{2}\|g\|_\mu^2 + \frac{\beta}{2}\|g - b\|_\mu^2 = \frac{1+\beta}{2}\|g - c_\beta b\|_\mu^2 + \frac{\beta}{2(1+\beta)}\|b\|_\mu^2.$$

The final term is independent of g. Therefore, with $u = g - c_\beta b$, problem (4) is equivalent to

$$\begin{aligned} u_\beta^\star = \arg\min_{u \in L^2(\mu;\mathbb{R}^d)} \quad & \frac{1}{2}\mathbb{E}_\mu\|u\|^2 \\ \text{subject to} \quad & -\nabla\cdot\big(\mu(c_\beta b + u)\big) = \lambda(\nu - \mu), \end{aligned}$$

and $g_\beta^\star = c_\beta b + u_\beta^\star$.

This equivalent formulation gives compensation a precise optimality criterion: among all modifications of the scaled prior $c_\beta b$ that complete the prescribed marginal response, it selects the one with minimum instantaneous correction energy. The prior is accounted for before the correction is fitted, including any insufficient or excessive response it supplies. At $\beta = 0$, the scaled prior vanishes and the problem reduces to the minimum-energy field for the prescribed density change.

The feasible set is still $g_0^\star + N_\mu$. Write $g = g_0^\star + n$ and $b = P_\nabla^\mu b + b_\perp$. Writing J_β for the objective of (4), its part that depends on n is

$$\frac{1}{2}\|n\|^2 + \frac{\beta}{2}\|n - b_\perp\|^2.$$

Its first-order condition is $(1 + \beta)n = \beta b_\perp$, proving (14). The Lagrangian with multiplier h is

$$\mathcal{L}(g, h) = \frac{1+\beta}{2}\|g\|^2 - \beta\langle g, b\rangle + \frac{\beta}{2}\|b\|^2 + \ell(h) - \langle\nabla h, g\rangle.$$

Its minimizer in g is $(\beta b + \nabla h)/(1 + \beta)$, so

$$\min_{g \text{ feasible}} J_\beta(g) = \frac{\beta}{2}\mathbb{E}_\mu\|b\|^2 + \max_h \mathcal{D}_\beta(h). \tag{24}$$

At the optimum, the two expressions for $g_\beta^\star$ give

$$\frac{\nabla h_\beta^\star}{1+\beta} = g_0^\star - \frac{\beta}{1+\beta} P_\nabla^\mu b.$$

Minimum-energy correction relative to the scaled prior. Under the same assumptions, fix $c_\beta = \beta/(1+\beta)$. Every correction u for which $c_\beta b + u$ satisfies the density constraint has the form

$$u = g_0^\star - c_\beta P_\nabla^\mu b + n, \qquad n \in N_\mu.$$

Indeed, every feasible total field belongs to $g_0^\star + N_\mu$, and the weighted divergence-free component of $c_\beta b$ can be absorbed into n.

The field $a_\beta = g_0^\star - c_\beta P_\nabla^\mu b$ belongs to G_μ. Since G_μ and N_μ are orthogonal,

$$\|u\|_{L^2(\mu)}^2 = \|a_\beta\|_{L^2(\mu)}^2 + \|n\|_{L^2(\mu)}^2.$$

Consequently, the unique minimum-energy correction is

$$u_\beta^\star = g_0^\star - c_\beta P_\nabla^\mu b,$$

and its total field is

$$c_\beta b + u_\beta^\star = g_0^\star + c_\beta b_\perp = g_\beta^\star.$$

This proves that the original flow-prior formulation and the minimum-correction formulation select the same field.

For the normalized readout $u_\beta[h]$ of Section 3.5, the optimality condition gives, for every admissible test function f,

$$\mathbb{E}_\mu\langle u_\beta^\star, \nabla f\rangle = \lambda(\mathbb{E}_\nu f - \mathbb{E}_\mu f) - c_\beta \mathbb{E}_\mu\langle b, \nabla f\rangle.$$

The effect of subsequent magnitude processing is analyzed in Appendix C.2.

C.1 Population fitting error and the weak residual

For fixed μ, ν, b, β, define $g_\beta[h] = (\beta b + \nabla h)/(1+\beta)$. The first-order condition at $h_\beta^\star$ is

$$\ell(f) = \frac{1}{1+\beta}\mathbb{E}_\mu\langle \nabla h_\beta^\star + \beta b, \nabla f\rangle.$$

Expanding the quadratic dual around this optimum cancels the linear terms and gives

$$\mathcal{D}_\beta(h_\beta^\star) - \mathcal{D}_\beta(h) = \frac{\|\nabla h - \nabla h_\beta^\star\|_{L^2(\mu)}^2}{2(1+\beta)} = \frac{1+\beta}{2}\|g_\beta[h] - g_\beta^\star\|_{L^2(\mu)}^2. \tag{25}$$

The identity controls population field error at fixed distributions; a minibatch loss does not measure the unknown population optimum.

For any $g \in L^2(\mu; \mathbb{R}^d)$, define its weak density residual by

$$\mathcal{R}_g(f) = \lambda(\mathbb{E}_\nu f - \mathbb{E}_\mu f) - \mathbb{E}_\mu\langle g, \nabla f\rangle. \tag{26}$$

Since $g_\beta^\star$ is feasible, $\mathcal{R}_g(f) = \mathbb{E}_\mu\langle g_\beta^\star - g, \nabla f\rangle$. Taking the supremum over $\|\nabla f\|_{L^2(\mu)} \le 1$ yields

$$\|\mathcal{R}_g\|_{\dot{H}^{-1}(\mu)} = \|P_\nabla^\mu(g - g_\beta^\star)\|_{L^2(\mu)}.$$

For $g = g_\beta[h]$, the field difference belongs to G_μ, so this equals the field error in (25). Independent test functions examine selected responses of this residual. The numerical target at $\beta = 9$ is $g_0^\star - 0.9 P_\nabla^\mu b$ for the correction $\nabla h_9/10$.

C.2 Effect of correction scaling

Let $c_\beta = \beta/(1+\beta)$ and $\mathcal{B}_\mu v = -\nabla\cdot(\mu v)$. From $g_\beta^\star = c_\beta b + \nabla h_\beta^\star/(1+\beta) = g_0^\star + c_\beta b_\perp$ one obtains

$$\frac{\nabla h_\beta^\star}{1+\beta} = g_0^\star - c_\beta P_\nabla^\mu b, \qquad \mathcal{B}_\mu \frac{\nabla h_\beta^\star}{1+\beta} = \lambda(\nu - \mu) - \mathcal{B}_\mu(c_\beta b). \tag{27}$$

The potential component therefore realizes the difference between the target density change and the change already produced by the base direction. For fixed μ, ν, b, a deterministic scalar $s \in [0,1]$, shared over the state space, gives

$$g_s^\star = c_\beta b + s\frac{\nabla h_\beta^\star}{1+\beta}, \tag{28}$$

$$\mathcal{B}_\mu g_s^\star = s\lambda(\nu - \mu) + (1-s)\mathcal{B}_\mu(c_\beta b). \tag{29}$$

This identity interpolates between the scaled base drift and exact relaxation. An empirical batch-RMS cap is random and does not directly equal this deterministic scaling. A spatially varying $s(x)$ also contributes a product-rule divergence term.

For example, if $\mu = \nu$ and b is a nonzero gradient field, the ideal $\beta = 9$ correction is $-0.9b$, whereas the 10% cap yields $-0.1b$ and total field $0.8b$. The cap therefore controls supplied magnitude while changing the exact residual response.

C.3 Complementarity in a restricted potential class

The base direction and the potential have different sources of approximation error. A reference-weighted error bound for a teacher does not automatically control its error on student queries: for an error field e and $\mu \ll q$,

$$\|e\|^2_{L^2(\mu)} \le \left\| \frac{\mathrm{d}\mu}{\mathrm{d}q} \right\|_\infty \|e\|^2_{L^2(q)}$$

gives a finite uniform bound when the density ratio is bounded. Mean-velocity integration additionally depends on the states visited along its path. Conversely, a finite potential class may fail to express all components of the desired weak response. The following fixed-distribution result makes a conditional form of complementarity precise.

Proposition 4 (Residual fitting in a fixed linear gradient class)**.** *Under Theorem 1, fix the laws μ, ν and a base field $b \in L^2(\mu; \mathbb{R}^d)$. Let $P = P^\mu_\nabla$, let $V \subseteq G_\mu$ be a closed linear space of admissible potential gradients, and let Π_V be its orthogonal projection. For finite $\beta \ge 0$, write $c = \beta/(1+\beta)$ and $a = g_0^\star - cPb$. Constrain the normalized potential gradient $v = \nabla h/(1+\beta)$ to V. The population dual is maximized at $v^\star_\beta = \Pi_V a$, and every $\widehat{v} \in V$ satisfies*

$$\|\mathcal{R}_{cb+\widehat{v}}\|^2_{\dot{H}^{-1}(\mu)} = \|(I - \Pi_V)a\|^2_{L^2(\mu)} + \|\widehat{v} - v^\star_\beta\|^2_{L^2(\mu)}. \tag{30}$$

Here $\mathcal{R}$ is the weak residual in (26)*. In particular, for the same scaled base and no magnitude operation, define*

$$g_{\mathrm{add}} = cb + \Pi_V g_0^\star, \qquad g_{\mathrm{res}} = cb + \Pi_V(g_0^\star - cPb).$$

Then

$$\|\mathcal{R}_{g_{\mathrm{add}}}\|^2 - \|\mathcal{R}_{g_{\mathrm{res}}}\|^2 = c^2\|\Pi_V Pb\|^2_{L^2(\mu)} \ge 0. \tag{31}$$

Proof. By the Riesz identity, $\lambda(\mathbb{E}_\nu h - \mathbb{E}_\mu h) = \langle g_0^\star, \nabla h\rangle_\mu$. Substituting $\nabla h = (1+\beta)v$ into the dual gives a constant minus $(1+\beta)\|v - a\|^2/2$. Maximizing over V therefore gives $v^\star_\beta = \Pi_V a$. The dual norm of the weak residual is $\|g_0^\star - Pg\|$; for $g = cb + \widehat{v}$, it is $\|a - \widehat{v}\|$. Orthogonality between V and $V^\perp$ proves (30). Finally,

$$P(g_{\mathrm{add}} - g_0^\star) = c\Pi_V Pb - (I - \Pi_V)a,$$

and these two terms are orthogonal. Subtracting the squared residual of g_{res} proves (31). □

Residual versus additive fitting with the same base. Equation (30) separates the response outside the potential class from the error in fitting the representable correction. The variable v in Proposition 4 is the normalized potential-gradient correction $u_\beta[h]$ in Section 3.5. Equation (31) compares two objectives that retain the same scaled base cb: the additive objective fits the full target response, whereas the residual objective first accounts for the response supplied by the base. With exact population fitting in V, the residual objective removes the additional error $c^2\|\Pi_V Pb\|^2_{L^2(\mu)}$ associated with adding that representable base response twice.

Flow-assisted residual fitting versus a pure potential. A separate comparison asks whether retaining the base helps at all. The best pure potential field in V is $g_{\mathrm{pot}} = \Pi_V g_0^\star$. Define

$$r = (I - \Pi_V)g_0^\star, \qquad z = (I - \Pi_V)Pb.$$

Here r is the target response unavailable to the potential, and z is the base response in that same subspace. The pure potential leaves residual r, while the optimally fitted flow-assisted field leaves residual $r - cz$. Hence

$$\|\mathcal{R}_{g_{\mathrm{pot}}}\|^2_{\dot{H}^{-1}(\mu)} - \|\mathcal{R}_{g_{\mathrm{res}}}\|^2_{\dot{H}^{-1}(\mu)} = 2c\langle r, z\rangle_{L^2(\mu)} - c^2\|z\|^2_{L^2(\mu)}.$$

Retaining the base strictly improves this error exactly when

$$2c\langle r, z\rangle_{L^2(\mu)} > c^2\|z\|^2_{L^2(\mu)}.$$

Thus the base complements the restricted potential when its otherwise unrepresentable response has a suitable direction and magnitude relative to the missing target response.

Scope. These statements concern fixed laws, a fixed closed linear gradient class, exact population fitting, and fields before magnitude processing. The full gradient space gives every exact $g^\star_\beta$ the same prescribed marginal response; when b is a gradient field, all these ideal total fields equal $g^\star_0$. The flow prior also leaves the solvability requirement $\nu - \mu \in \dot{H}^{-1}(\mu)$ unchanged. The nonlinear CNN, finite fitting budget, magnitude rule, and parameterized student used in the experiments introduce additional approximation and optimization effects. Their endpoint quality is assessed separately.

D Reference distributions

D.1 Independent affine Gaussian noising

For endpoint laws P, Q, define

$$p_t = (a_t I)_\# P * \mathcal{N}(0, \sigma_t^2 I), \qquad q_t = (a_t I)_\# Q * \mathcal{N}(0, \sigma_t^2 I).$$

When $a_t \neq 0$, $p_t = q_t$ if and only if $P = Q$. Indeed,

$$\varphi_{p_t}(\xi) = \varphi_P(a_t\xi)\exp(-\sigma_t^2\|\xi\|^2/2),$$

and the Gaussian characteristic function never vanishes. The same identity for q_t and the invertibility of scaling by a_t prove the result. A coupling using shared Gaussian noise also gives

$$W_2(p_t, q_t) \leq |a_t| W_2(P, Q).$$

There is no general stable reverse inequality. Exact identifiability therefore does not establish stable endpoint recovery from finite noised samples.

ImageNet uses independent affine noising on both sides. SANA takes the reference directly from a teacher trajectory, so that construction does not instantiate this identifiability statement. It still uses the dual of Section 3 on the query laws actually compared.

D.2 SANA query constructions

SANA keeps independent re-noising only on the student side. The student query is $\mathcal{N}_t(P_\theta)$; the reference is the teacher trajectory ν_t^T, not independently re-noised teacher endpoints. Fix a prompt and write $\mathcal{N}_t(P) = \mathrm{Law}((1-t)X + t\xi)$ for $X \sim P$ independent of $\xi \sim \mathcal{N}(0, I)$. With student and teacher endpoint laws P_θ, P_T, the difference SANA fits is

$$\mathcal{N}_t(P_\theta) - \nu_t^T = \big[\mathcal{N}_t(P_\theta) - \mathcal{N}_t(P_T)\big] + \big[\mathcal{N}_t(P_T) - \nu_t^T\big]. \tag{32}$$

The first term is the independently noised endpoint gap and is zero once $P_\theta = P_T$. The second is the gap between re-noising the teacher endpoint and sampling the teacher trajectory at time t. With CFG and Euler integration, that second term can remain after the student matches the teacher endpoint, so SANA fits generation error together with reference-construction error.

D.3 Trajectory versus endpoint re-noising

Table 2 uses the teacher trajectory ν_t^T. We compare it with teacher-endpoint re-noising $\mathcal{N}_t(P_T)$ in two ways. On 64 prompts, at each of $t = 0.2, 0.5, 0.8$, 16 re-noised teacher-endpoint samples are compared with 16 independent trajectory samples, CFG 4.5, and Euler step bound 0.03. Unbiased multi-scale RBF MMD^2 has 95% prompt-bootstrap intervals, scaled by 1000, of $[-0.732, 0.731]$, $[-0.211, 0.751]$, and $[-0.048, 0.136]$, using 4000 resamples. All three intervals span zero: sixteen samples per prompt neither separate the two teacher queries nor establish that they share a law.

This separate pair is trained on a shared fixed pool of 3000 prompts, distinct from the training-text selection used in Table 2. At 3000 steps, endpoint re-noising yields FID 25.359, compared with 27.930 for the paired trajectory-reference run. Four of six scores improve; PickScore and CLIPScore are slightly lower. Both runs complete 272 student and 2728 auxiliary updates; endpoint re-noising records longer training time. Evaluation uses fixed 5k prompts, NFE 4 and CFG 1 (Appendix E.5).

E IMPLEMENTATION AND EVALUATION

The algorithm changes neither the auxiliary objective nor the model-update schedule. During potential optimization, P, Q, b_P are detached from their generating models, but the gradient-energy term is differentiated through $\nabla_x h_\omega$ with respect to ω. During student optimization, the only backward path is

$$\theta \longrightarrow X_t^\theta \longrightarrow \mathcal{L}_{\text{student}}.$$

Teacher and auxiliary predictions, potential gradients, and magnitude operations all supply stopped-gradient directions. Potential updates are counted separately from the global student/auxiliary steps.

E.1 FROM A TRAINING DIRECTION TO QUERY DISPLACEMENT

For the B_{s} queries retained for student backpropagation, hold the sampled inputs, times, conditions, and directions fixed. Write $J_i = \partial X_i^\theta / \partial \theta$, and suppose the query map is twice differentiable near θ. One unclipped SGD step on (13) gives

$$\Delta\theta = \frac{\eta}{B_{\text{s}} d} \sum_{j=1}^{B_{\text{s}}} J_j^\top g_j^{\text{impl}}, \qquad X_i^{\theta+\Delta\theta} - X_i^\theta = \frac{\eta}{B_{\text{s}} d} \sum_{j=1}^{B_{\text{s}}} J_i J_j^\top g_j^{\text{impl}} + O(\eta^2). \tag{33}$$

The Jacobian factors couple the queries and transform both direction and magnitude. The experiments additionally use AdamW, with moment history, parameter-dependent scaling, and gradient clipping. A bound on correction RMS in query coordinates is therefore a bound on the supplied direction, not on the realized query or parameter displacement.

E.2 POTENTIAL ARCHITECTURE AND OPTIMIZATION

Both high-dimensional experiments use a conditional CNN. The latent query is concatenated with four spatial channels: horizontal and vertical coordinates and their sine values. A 3×3 convolution produces 64 channels. Four residual blocks output 64, 96, 128, and 192 channels, with downsampling in the last three blocks. Each has two 3×3 convolutions, GroupNorm, SiLU, and a residual connection. Frozen class or pooled text features are projected to 128 dimensions and combined with Fourier encodings of t, r to condition the blocks. Global average pooling and a $192 \to 192 \to 1$ head produce the scalar potential.

The potential uses FP32 AdamW at learning rate 10^{-4}, gradient-norm clipping at 1, 20 warm-up updates, and five updates before every student update. EMA is updated once per student update with $\alpha = 2^{-1/8}$. The common relaxation rate is $\lambda = 1$. The entire loss in (10), including its mean penalty, is multiplied by $s_{\text{loss}} = 1$ for ImageNet and $1/32768$ for SANA; the SANA factor is a common scale on all terms.

On ImageNet the online potential forward uses BF16 autocast; parameters, optimizer states, and loss arithmetic remain FP32. The EMA potential is evaluated in FP32. The mean penalty is computed within each four-query local batch before gradients are averaged across ranks. SANA potential fitting uses FP32 and the microbatch accumulation below.

E.3 FINITE-SAMPLE POTENTIAL FITTING

The population result requires a bounded linear marginal functional on the weighted Sobolev space. For disjoint finite sets P, Q, an unrestricted smooth class can instead interpolate $h_A = -A$ near P and $h_A = A$ near Q, with zero gradients on P. Then

$$\widehat{\mathcal{D}}_\beta(h_A) = 2\lambda A - \frac{\beta^2}{2(1+\beta)} \widehat{\mathbb{E}}_P \|b_P\|^2, \qquad \widehat{m}_{h_A} = 0,$$

which grows without bound. The mean penalty fixes the additive offset. The implementation restricts fitting through the finite CNN, weight decay, five optimization steps per student update, and renewed queries; clipping and EMA control the parameter updates.

E.4 IMAGENET SAMPLE CONSTRUCTION AND TRAINING

The CNN input latent is $4 \times 32 \times 32$; its class feature comes from the frozen SiT class embedding. Teacher, student, and auxiliary share the pretrained multistep model initialization, with separate rank/alpha $64/64$ student and auxiliary LoRAs targeting `attn.qkv` in the transformer blocks. Student and auxiliary use AdamW with learning rate 10^{-4}, momenta $(0.9, 0.95)$, weight decay zero, and $\epsilon = 10^{-8}$. Potential AdamW uses $(0.9, 0.999)$, weight decay 10^{-4}, and $\epsilon = 10^{-8}$.

For each student update, independent $U, V \sim \mathrm{Uniform}(0, 1)$ determine one shared time pair,

$$t = 10^{-6} + (1 - 2 \cdot 10^{-6})U, \qquad r = t\big[10^{-6} + (1 - 10^{-6})V\big].$$

The pair is shared across the batch. Auxiliary updates instead use an independent uniform time per query. Mean velocities integrate from t to r by Euler steps $\min(0.03, t_{\mathrm{current}} - r)$. Four-step student endpoint generation and evaluation use the grid $1, 3/4, 1/2, 1/4, 0$, with velocity evaluations at the first four points.

Student fitting queries and actual student-update queries share class and time context but are independently generated. The five potential steps reuse the fitting student queries and resample reference noise. An additional independent reference sample can evaluate the potential objective away from the fitting sample. Base directions are evaluated at the queries to which they apply. Common class, data-index, and time random streams do not depend on the experimental arm, so the additional potential operations do not alter the scheduled training inputs. Different evolving students still generate different query values.

The global batch of 32 spans eight GPU processes and contains two classes with 16 queries each. RMS is computed over the global batch and all latent coordinates. All models complete 560 student and 1440 auxiliary updates; FWCD (β=9) and Additive $\mathcal{D}_0$ add $20 + 5(560) = 2820$ potential updates. The R1–R3 $M = 1$ comparison in Table 13 uses $20 + 1(560) = 580$ potential updates with the same student and auxiliary counts.

Figure 3 uses the R2 checkpoints at step 2000, four Euler steps and the 10% cap protocol. Columns (a)–(h) correspond to evaluation indices 000000, 007143, 014286, 021429, 028571, 035714, 042857 and 049999, evenly spaced over the 50k list. Rows are MFD, Attenuation, FWCD-Shuffle and FWCD, with shared class and initial noise. Full-evaluation FID values are reported in Table 7.

E.5 SANA TRAINING AND QUERY REUSE

The frozen pretrained model is `https://huggingface.co/Efficient-Large-Model/Sana_1600M_1024px_BF16_diffusers`. Student and auxiliary each train a rank/alpha $64/64$ LoRA on `to_q`, `to_k`, and `to_v`; the teacher has no LoRA. Training uses BF16, batch size one, and 8-bit AdamW with learning rate 10^{-4}, weight decay 10^{-4}, and gradient-norm cap 0.5. The auxiliary retains the MFD mean-flow objective. The student retains the query construction, mean-velocity estimator, and update schedule, and uses the implemented FWCD field in the stopped-gradient surrogate (13). Ten auxiliary updates alternate with one student update over 3000 steps, giving 272 student and 2728 auxiliary updates. The time pair (t, r) is the maximum and minimum of two independent logit-normal draws $\mathrm{sigmoid}(Z - 0.4)$, $Z \sim \mathcal{N}(0, 1)$, rounded to 0.001; when rounding gives $t = r$, the instantaneous velocity is used. Mean-velocity and teacher-reference Euler integration follow the official MFD SANA implementation (Zhao et al., 2026) with nominal spacing 0.03 and teacher CFG 4.5. Four-step student generation uses the multistep DPM-Solver++ scheduler of Diffusers 0.36.0 (order 2, flow shift 3) with timestep indices $(999, 899, 749, 499)$; these sampling levels are separate from the training integration grid.

The potential input has 32768 latent coordinates, and text conditioning is the mean of valid token features. For each student update, 16 student queries and independent reference queries share one prompt and time context. The five potential optimization steps reuse that set with gradient accumulation over microbatches of four. The marginal-fitting batch is separated from the batch retained for generator backpropagation: fitting uses 16 queries, while one retains the student computation graph. This estimator is not equated with backpropagating through all 16. Compensation and its magnitude rule are computed for all 16 queries; only the first query backpropagates to the student parameters, and the potential receives $20 + 5(272) = 1380$ updates. The $M = 1$ and $M = 10$ comparisons in Table 14 use $20 + 1(272) = 292$ and $20 + 10(272) = 2740$ potential updates with the

same student and auxiliary counts. FWCD-Shuffle uses a permutation without fixed points within the 16-query set, restores the reassigned correction's original RMS, and rescales the remaining 15 vectors to preserve total correction energy; capped shuffling applies the cap first. These constraints compare the operation before and after permutation within the same arm and do not equate magnitudes across the independently evolving FWCD and FWCD-Shuffle trajectories.

Prompts come from a fixed random permutation of 479,410 cleaned texts: model updates use the first 3000, potential warm-up draws 20 further prompts from the full pool, and 5000 evaluation prompts are taken from entry 10001 onward after deduplication and exclusion, with no overlap with model-update or warm-up texts. Evaluation uses a fixed prompt order and initial noise shared across methods, with image seed $s_{\mathrm{S1}} + i$ (Table 5) for prompt i. Every evaluated student uses checkpoint 3000, NFE 4 and CFG 1; evaluation reference images use the teacher at NFE 20, CFG 4.5.

Figure 5 uses the natural-magnitude experiment at checkpoint 3000, NFE 4 and CFG 1. Rows (a)–(e) select indices 0, 1000, 2000, 3000 and 4000, evenly spaced in the fixed 5k list. Columns are MFD, Attenuation, FWCD-Shuffle and FWCD. The short row labels refer to these complete prompts:

(a) nature, tree, and forest image
(b) 9. Lennart Jirlow, In the park.
(c) ice queen: Beautiful young woman in crown and silver top standing near frozen mirror. Snow queen. Copy space.
(d) Spread A Little Happiness by Bob Barker
(e) "Barn and Goldenrod" - 22"x 28" - Oil

E.6 Evaluation scores

FID uses Inception-v3 features (Heusel et al., 2017). ImageNet KID uses the unbiased polynomial-kernel MMD estimate in the same feature space (Bińkowski et al., 2018), with displayed values multiplied by 1000.

SANA aesthetic scores use the LAION-Aesthetics Predictor V1. Its linear head `sa_0_4_vit_l_14_linear.pth` receives normalized 768-dimensional CLIP ViT-L/14 image features.[1] ImageReward uses the ImageReward-v1.0 checkpoint and its configuration via the official `image-reward` implementation.[2] HPSv2 uses checkpoint version 2.1[3]. PickScore uses PickScore-v1 and reports image–text embedding cosine, not a two-image selection probability.[4] CLIPScore uses CLIP ViT-B/32 and the score $100 \max\{\cos(\mathrm{image}, \mathrm{text}), 0\}$. Each of these five scores is an arithmetic mean over the same 5000 generated images. SANA's distribution references are teacher images, while ImageNet uses real-image reference statistics.

Figure 1(a) uses 50k ImageNet samples from R1 with 560 student updates per method; DMD2, VSD, and CD are SiT-LoRA adaptations with method-specific auxiliary budgets. Panel (b) uses the scores in Table 2, min–max normalized within each metric over the seven displayed methods, with FID reversed. FWCD uses the default configuration for each task.

F Randomization

Table 5 defines the replicate IDs used throughout the paper. Methods within a comparison share initialization and the evaluation stream. ImageNet uses R1–R6; SANA uses S1. Write s_{S1} for its tabulated seed value. SANA image i uses $s_{\mathrm{S1}} + i$; the COCO prompt permutation uses a separate generator initialized with s_{S1}.

[1] `https://github.com/LAION-AI/aesthetic-predictor`
[2] `https://github.com/zai-org/ImageReward`
[3] `https://github.com/tgxs002/HPSv2`
[4] `https://huggingface.co/yuvalkirstain/PickScore_v1`

Table 5: Random settings. Per-run results are indexed by replicate ID; s_{S1} is the seed value assigned to S1.

Comparison	Replicate	Training RNG	Evaluation RNG
Default and budget	R1	20260830	20260616
Default and budget	R2	20260903	20260616
Default and budget	R3	20260904	20260616
Relative-magnitude replay	R4	20260901	20260616
Relative-magnitude replay	R5	20260902	20260616
Pairing displacement	R6	20260828	20260616
SANA	S1	42	$s_{\mathrm{S1}} + i$

G IMAGENET PER-RUN QUALITY RESULTS

G.1 SAME-SEED BASELINE COMPARISON

Table 6 compares the R1 models used in Figure 1(a). DMD2, VSD (Wang et al., 2023), and CD are SiT-LoRA adaptations. All methods use 560 student updates, the same evaluation seed, and 50k four-step samples with CFG 1. Auxiliary update counts are 2800 for DMD2, 560 for VSD, zero for CD, and 1440 for MFD/FWCD; DMD2 also uses 2800 discriminator updates, and CD selects EMA student weights. This comparison matches student updates rather than total computation.

Table 6: ImageNet-256 baseline comparison on R1. All entries are single-training-run estimates.

Method	FID ↓	KID ($\times 1000$) ↓
DMD2	8.760	3.477
VSD	6.558	1.436
CD	24.734	17.192
MFD	4.457	0.632
FWCD (β=9)	**4.233**	**0.577**

G.2 MATCHED REPLICATES AND EVALUATION UNCERTAINTY

Table 1 reports means and between-run standard deviations (denominator $n - 1$) over the three runs in Table 7. Improvement percentages use unrounded measurements. Both aligned potential arms have lower FID and KID than MFD in each replicate. The ordering of MFD and Attenuation varies across replicates.

Early-only compensation improves both means over MFD, while full FWCD (β=9) improves them further. In R3, full compensation also outperforms early-only on both metrics, supporting continued compensation beyond the first-quarter window.

Table 9 uses the R1 models from Table 7, checkpoint 2000 and 50k generated samples. FWCD and FWCD-Shuffle are the capped $\mathcal{D}_\beta(\beta{=}9)$ arms. The paired intervals quantify evaluation-sample variation conditional on these trained models, not variation across training runs.

All three FID intervals in Table 9 lie below zero. For KID, methods share generated-image indices and reference subsets, and groupwise differences form the paired estimate and interval. This estimator is separate from subtracting independently reported full-table KID values. Its intervals cross zero.

H SANA RESULTS AND CORRECTION MAGNITUDE

Figure 5 compares the natural-magnitude arms on shared prompts; their scores are reported in Table 2.

On SANA, the 10% cap activates on 39.34% of FWCD (cap) student updates and 28.31% of FWCD-Shuffle (cap) updates. The correction-to-base RMS ratio on the query used for the student update averages 5.56% and 5.29%, respectively.

Table 7: ImageNet-256 per-run results for the method comparison (10% cap), with 50k images at checkpoint 2000. Additive $\mathcal{D}_0$ fits $\mathcal{D}_0$ with the same $0.9b$ mixture as FWCD (β=9). The additive $\mathcal{D}_0$-Shuffle rows complete the comparison in Appendix I.4.

Rep.	Method	FID ↓	KID ×1000 ↓
R1	MFD	4.457	0.632
	Attenuation	4.413	0.640
	Additive $\mathcal{D}_0$	**3.871**	**0.537**
	FWCD (β=9)	4.233	0.577
	FWCD-Shuffle	4.286	0.587
	Additive $\mathcal{D}_0$-Shuffle	4.200	0.573
R2	MFD	4.748	0.725
	Attenuation	4.822	0.735
	Additive $\mathcal{D}_0$	**4.147**	**0.558**
	FWCD (β=9)	4.322	0.613
	FWCD-Shuffle	4.867	0.722
	Additive $\mathcal{D}_0$-Shuffle	4.528	0.628
R3	MFD	4.448	0.632
	Attenuation	4.411	0.829
	Additive $\mathcal{D}_0$	4.160	**0.529**
	FWCD (β=9)	**4.112**	0.585
	FWCD-Shuffle	4.173	0.540
	Additive $\mathcal{D}_0$-Shuffle	4.223	0.662

Table 8: Early-only compensation control (10% cap). FWCD (β=9) applies compensation for the first 140 student updates (one quarter of 560, corresponding to global step 490) and then stops potential fitting and EMA, returning the student direction to b; the potential is updated $20 + 5 \times 140 = 720$ times. Remaining protocol and replicates R1–R3 match Table 7; checkpoint 2000, 50k images. MFD and full-FWCD (β=9) columns are taken from Table 7.

Rep.	Method	FID ↓	KID ×1000 ↓
R1	MFD	4.457	.632
	FWCD (β=9), full	4.233	**.577**
	FWCD (β=9), first 140	**4.232**	.603
R2	MFD	4.748	.725
	FWCD (β=9), full	**4.322**	.613
	FWCD (β=9), first 140	4.330	**.599**
R3	MFD	4.448	.632
	FWCD (β=9), full	**4.112**	**.585**
	FWCD (β=9), first 140	4.375	.628
Mean ± std	MFD	4.551 ± 0.170	0.663 ± 0.054
	FWCD (β=9), full	$\mathbf{4.222} \pm 0.105$	$\mathbf{0.592} \pm 0.019$
	FWCD (β=9), first 140	4.313 ± 0.073	0.610 ± 0.016

In Figure 7, the natural correction-to-base RMS ratio stays near zero for roughly the first 50 student updates, then grows to 20–40% late in training with short spikes. Removing the group-mean direction leaves a fraction of approximately 0.5–0.9, indicating a substantial query-specific component. Potential parameter-gradient clipping occurs intermittently in the later stage.

Table 9: Paired 95% evaluation intervals for R1. Differences are FWCD minus the comparator; negative values favor FWCD.

Metric	Comparator	Difference	95% interval
FID	MFD	-0.224	$[-0.268, -0.180]$
	Attenuation	-0.180	$[-0.221, -0.140]$
	FWCD-Shuffle	-0.053	$[-0.093, -0.013]$
KID ($\times 1000$)	MFD	-0.052	$[-0.190, 0.086]$
	Attenuation	-0.056	$[-0.158, 0.048]$
	FWCD-Shuffle	-0.004	$[-0.131, 0.103]$

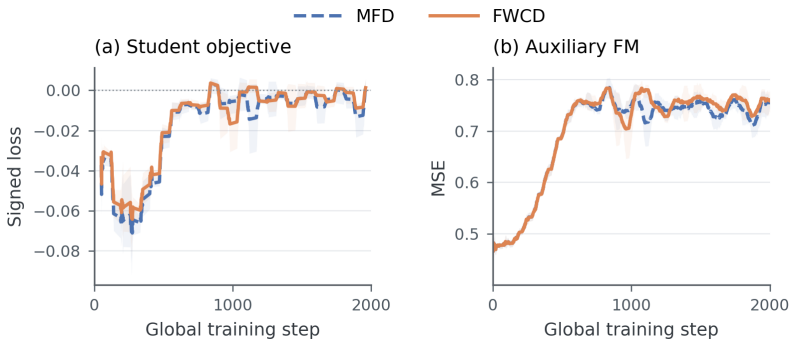

Figure 4: ImageNet-256 training curves for replicate R2. Left: signed student surrogate loss. Right: auxiliary Flow Matching MSE. Curves are backward moving averages. The signed student loss is not a sample-quality metric; both runs train to the end without divergence.

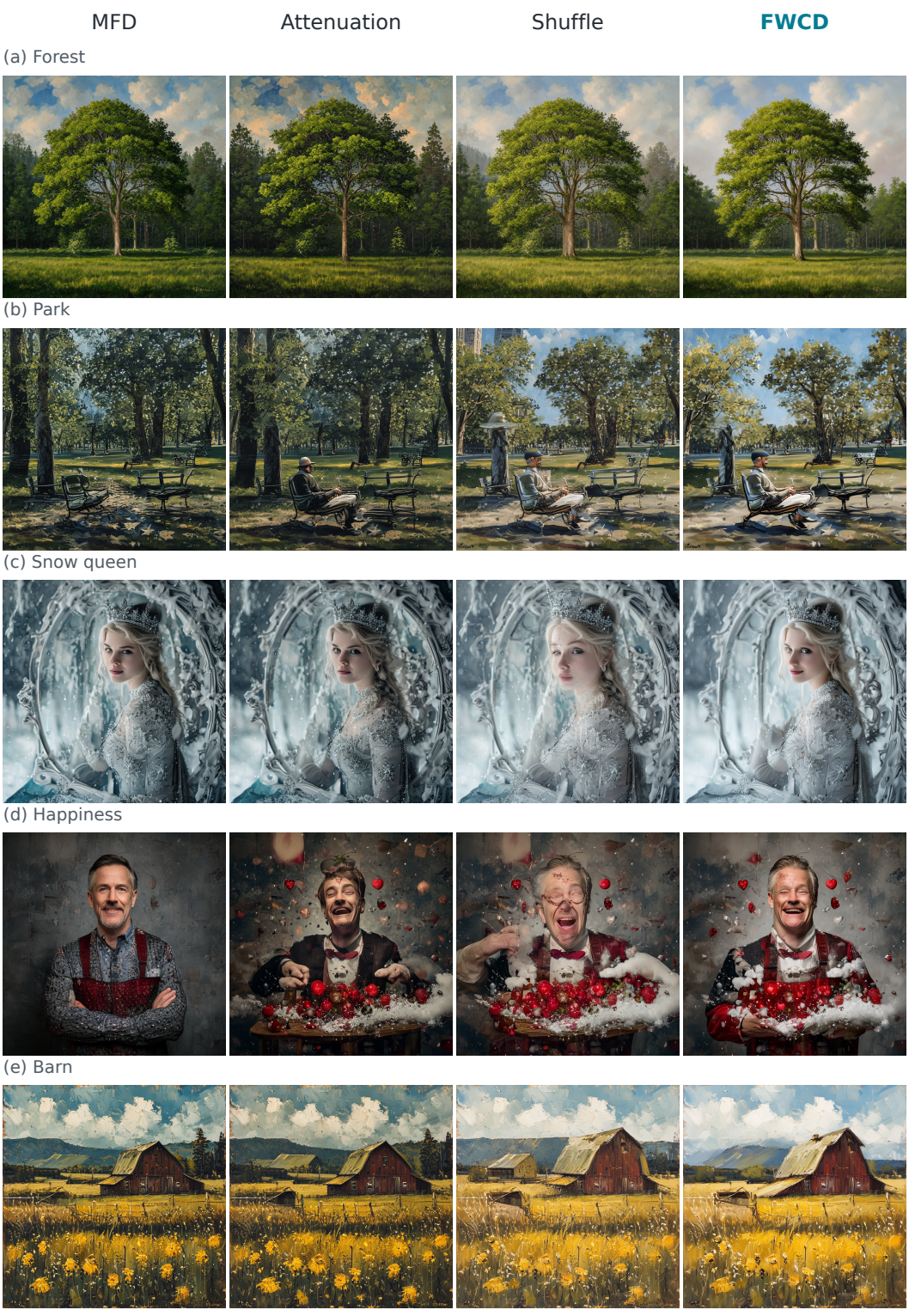

Figure 5: Four-step SANA samples for five shared prompts. Columns use matched initial noise; Shuffle denotes FWCD-Shuffle. Full prompts are listed in Appendix E.5.

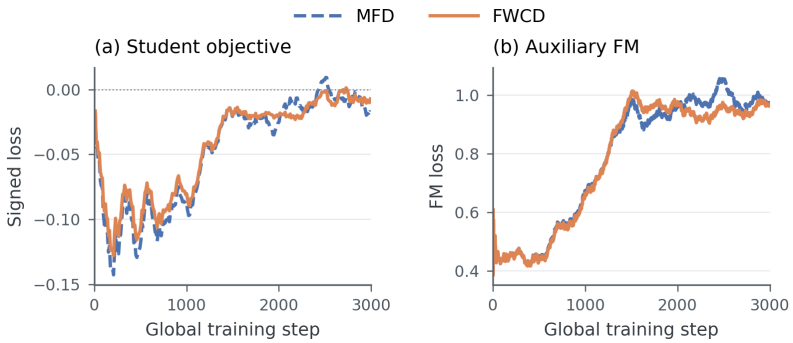

Figure 6: SANA-1.6B training curves. Left: signed distillation objective over 272 student updates. Right: auxiliary Flow Matching loss over 2728 auxiliary updates. Both are moving averages. The curves stay in a common range with no collapse; the signed student objective is not used to rank generation quality.

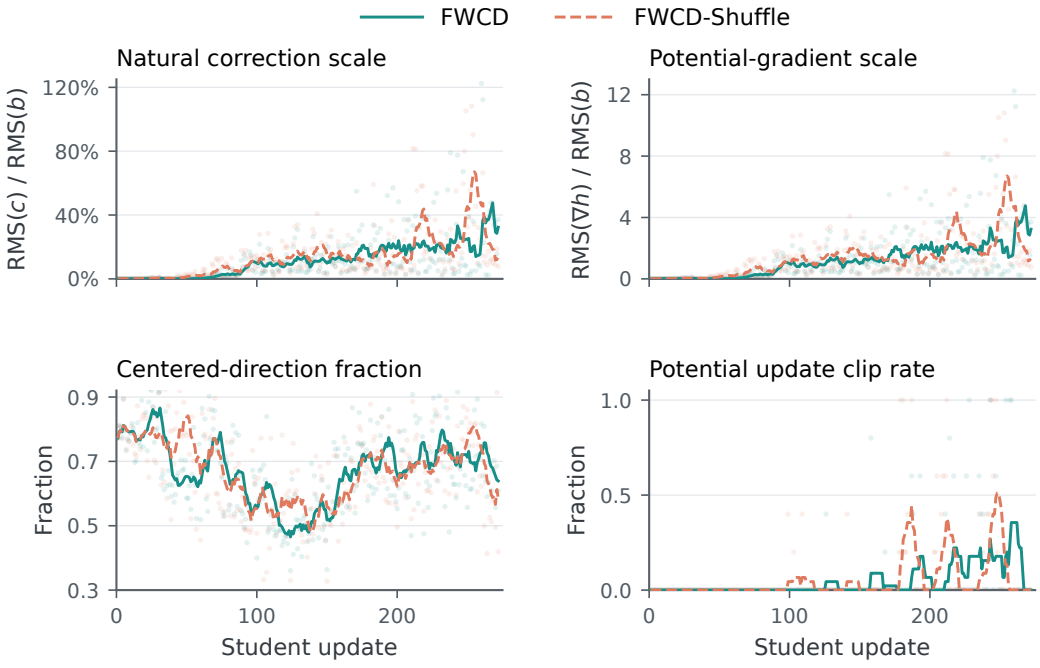

Figure 7: Natural-magnitude diagnostics for SANA FWCD and FWCD-Shuffle, as 9-update moving means. Top left: correction-to-base RMS ratio. Top right: $\mathrm{RMS}(\nabla h)/\,\mathrm{RMS}(b)$. Bottom left: fraction of the correction remaining after removing the group-mean direction. Bottom right: potential parameter-gradient clip rate.

I Potential objective and direction ablations

I.1 Definitions and relative magnitude (replay)

Arms are labeled by β and magnitude rule: Additive $\mathcal{D}_0$ (replay) fits the basic weak dual $\mathcal{D}_0$, and FWCD(β=9, replay) fits the prior dual $\mathcal{D}_\beta(\beta{=}9)$; Additive $\mathcal{D}_0$-Shuffle (replay) is the within-class shuffle of the former, and FWCD(β=9, replay, no-ν) is the latter with the marginal contrast removed. This section uses replayed relative magnitude, not the 10% cap of Table 1.

In the matched-relative-magnitude ablation we replay a common correction-to-base RMS sequence in each potential-ablation arm. At student update k, a candidate direction v is normalized as

$$\mathcal{N}_k(v; b_k) = r_k \,\mathrm{RMS}(b_k)\frac{v}{\mathrm{RMS}(v)}. \tag{34}$$

Table 10: Potential-objective comparison (replay). Checkpoint 2000, 50k. Mean of R4 and R5. Per-run values and paired differences are in Appendix I.

Method	FID ↓	KID ×1000 ↓
MFD	4.125	0.601
Attenuation	4.234	0.576
Additive $\mathcal{D}_0$ (replay)	**3.842**	**0.531**
Additive $\mathcal{D}_0$-Shuffle (replay)	4.195	0.709
FWCD(β=9, replay)	4.000	0.592
FWCD(β=9, replay, no-ν)	4.389	0.688

The six ImageNet directions at student update k are

$$\text{MFD: } g_k = b_k, \qquad \text{Attenuation: } g_k = 0.9b_k, \tag{35}$$

$$\text{Additive } \mathcal{D}_0 \text{ (replay): } g_k = 0.9b_k + \mathcal{N}_k(\nabla h_{0,\mathrm{EMA}}; b_k), \tag{36}$$

$$\text{Additive } \mathcal{D}_0\text{-Shuffle (replay): } g_k = 0.9b_k + \Pi_{\pi_{c,k}}\mathcal{N}_k(\nabla h_{0,\mathrm{EMA}}; b_k), \tag{37}$$

$$\text{FWCD}(\beta{=}9, \text{ replay}): \; g_k = 0.9b_k + \mathcal{N}_k(\nabla h_{9,\mathrm{EMA}}/10; b_k), \tag{38}$$

$$\text{FWCD}(\beta{=}9, \text{ replay, no-}\nu): \; g_k = 0.9b_k + \mathcal{N}_k(\nabla h^{\text{no-}\nu}_{9,\mathrm{EMA}}/10; b_k). \tag{39}$$

The no-ν potential separately optimizes the energy-only objective

$$\mathcal{D}^{\text{no-}\nu}_\beta(h)\big|_{\beta=9} = -\frac{1}{20}\mathbb{E}_\mu\|\nabla h + 9b\|^2$$

with the same mean penalty, architecture, update count, EMA, and magnitude sequence as FWCD(β=9, replay). The shuffled arm uses a within-class Sattolo permutation without fixed points after normalization. Each potential is fitted on its own arm's student distribution.

Additive $\mathcal{D}_0$ (replay) retains the same $0.9b$ base term as FWCD(β=9, replay). It tests the basic potential as an additive correction; it is not the pure constrained $\beta = 0$ total field $g_0^\star$. Before magnitude replay, the ideal no-ν correction is $-0.9P^\mu_\nabla b$, so its total field is $0.9b_\perp$. Removing marginal contrast therefore changes the target dynamics rather than producing a neutral capacity-only control. Ordinary within-class shuffling preserves the class-mean correction as well as its RMS, so it removes the current sample assignment without removing all distributional information.

The common ratio sequence r_k is taken from the R1 FWCD (β=9) run, indexed by student updates $k = 1, \ldots, 560$. It has mean approximately 0.083 and median approximately 0.100. For the four potential arms in the relative-magnitude comparison, the maximum relative discrepancy in $\mathrm{RMS}(\delta_k) = r_k\,\mathrm{RMS}(b_k)$ is approximately 6×10^{-8}. The mean correction–base cosines, first averaged over 560 updates within each run and then across runs, are 0.121 for Additive $\mathcal{D}_0$ (replay), 0.010 for Additive $\mathcal{D}_0$-Shuffle (replay), -0.023 for FWCD(β=9, replay), and -0.271 for FWCD(β=9, replay, no-ν). Equal relative magnitude therefore leaves different directions.

Table 10 reports the mean endpoint metrics from unrounded measurements over replicates R4 and R5; Table 11 lists the per-run values. Additive $\mathcal{D}_0$ (replay) is below MFD, and FWCD(β=9, replay)

below FWCD(β=9, replay, no-ν), on both metrics in both replicates. Attenuation is unstable across R4 and R5 (FID 4.619 versus 3.848), which is why its two-run mean in Table 10 is above MFD while it is below MFD in R5.

Table 11: Relative-magnitude objective comparison per training run (replay). Checkpoint 2000, 50k images. Bold marks the lowest value within each replicate.

Rep.	Method	FID ↓	KID ×1000 ↓
R4	MFD	4.138	0.603
	Attenuation	4.619	0.654
	Additive $\mathcal{D}_0$ (replay)	**3.903**	**0.503**
	Additive $\mathcal{D}_0$-Shuffle (replay)	4.051	0.553
	FWCD(β=9, replay)	4.117	0.586
	FWCD(β=9, replay, no-ν)	4.490	0.667
R5	MFD	4.113	0.599
	Attenuation	3.848	**0.497**
	Additive $\mathcal{D}_0$ (replay)	**3.780**	0.560
	Additive $\mathcal{D}_0$-Shuffle (replay)	4.339	0.866
	FWCD(β=9, replay)	3.883	0.599
	FWCD(β=9, replay, no-ν)	4.288	0.709

Table 12: Paired evaluation intervals for the matched-relative-magnitude six-arm experiment (replay). Each cell is a point difference and a 95% interval conditional on the trained models. Negative differences favor the left-hand method.

Rep.	Left minus right	FID	KID ×1000
R4	Additive $\mathcal{D}_0$ (replay) − MFD	$-.234$ $[-.269, -.202]$	$-.100$ $[-.112, -.082]$
	Additive $\mathcal{D}_0$ (replay) − Attenuation	$-.716$ $[-.766, -.668]$	$-.152$ $[-.173, -.107]$
	Additive $\mathcal{D}_0$ (replay) − Additive $\mathcal{D}_0$-Shuffle (replay)	$-.148$ $[-.184, -.112]$	$-.050$ $[-.067, -.041]$
	FWCD(β=9, replay) − no-ν	$-.373$ $[-.420, -.326]$	$-.082$ $[-.093, -.041]$
R5	Additive $\mathcal{D}_0$ (replay) − MFD	$-.333$ $[-.380, -.285]$	$-.039$ $[-.048, -.008]$
	Additive $\mathcal{D}_0$ (replay) − Attenuation	$-.068$ $[-.112, -.025]$	$+.063$ $[+.048, +.105]$
	Additive $\mathcal{D}_0$ (replay) − Additive $\mathcal{D}_0$-Shuffle (replay)	$-.558$ $[-.617, -.499]$	$-.306$ $[-.336, -.264]$
	FWCD(β=9, replay) − no-ν	$-.404$ $[-.454, -.357]$	$-.110$ $[-.126, -.087]$

For both metrics, the paired intervals for FWCD(β=9, replay) minus no-ν and for Additive $\mathcal{D}_0$ (replay) minus MFD lie below zero. Comparisons against attenuation and shuffling have metric-specific exceptions.

FID uses a paired influence-function bootstrap on the ordered generated features, with 2000 bootstrap resamples. KID uses 50 disjoint blocks of 1000 images, forms paired block differences, and bootstraps their mean. These intervals quantify evaluation variation conditional on each set of trained models and fixed real-reference statistics; they are not intervals for the mean across training runs.

I.2 CAP-PROTOCOL POTENTIAL OBJECTIVE

The $M = 5$ objective comparison uses the shared $0.9b$ mixture and 10% cap defined in Section 7.3. Table 7 reports its per-run values, and Table 1 reports the means. Additive $\mathcal{D}_0$ has lower FID and KID than $\mathcal{D}_\beta(\beta{=}9)$ in R1 and R2; R3 favors $\mathcal{D}_\beta(\beta{=}9)$ in FID. The budget comparison below tests whether this ordering persists with fewer potential updates.

I.3 POTENTIAL UPDATES PER STUDENT STEP

We vary the number M of potential updates around the default $M = 5$ in Algorithm 1. Each inner loop fits the current (μ_t, ν_t, b); the experiments test whether additional fitting improves the direction supplied to the student.

ImageNet. We compare $M = 1$ and $M = 5$ for both potential objectives under the protocol of Appendix E.4, retaining the 10% cap, eight GPUs, and 560 student updates. R1–R3 use the same 50k-image, four-step evaluation as Table 7, which also supplies the MFD and $M = 5$ results. Reducing M changes the potential-update count from $20 + 5 \times 560 = 2820$ to $20 + 560 = 580$.

Table 13: ImageNet potential-update count (10% cap). Checkpoint 2000, 50k images. M is the number of potential updates before each student update. Replicates match Table 7. Lower is better.

Rep.	Method	FID ↓	KID ×1000 ↓
R1	MFD	4.457	.632
	FWCD (β=9), $M = 5$	4.233	.577
	FWCD (β=9), $M = 1$	**4.005**	**.534**
	Additive $\mathcal{D}_0$, $M = 1$	4.011	.666
R2	MFD	4.748	.725
	FWCD (β=9), $M = 5$	4.322	.613
	FWCD (β=9), $M = 1$	**4.047**	**.505**
	Additive $\mathcal{D}_0$, $M = 1$	4.456	.724
R3	MFD	4.448	.632
	FWCD (β=9), $M = 5$	4.112	**.585**
	FWCD (β=9), $M = 1$	4.132	.641
	Additive $\mathcal{D}_0$, $M = 1$	**3.934**	.806
Mean ± std	MFD	4.551 ± 0.170	0.663 ± 0.054
	FWCD (β=9), $M = 5$	4.222 ± 0.105	0.592 ± 0.019
	FWCD (β=9), $M = 1$	$\mathbf{4.061} \pm 0.065$	$\mathbf{0.560} \pm 0.072$
	Additive $\mathcal{D}_0$, $M = 1$	4.134 ± 0.282	0.732 ± 0.071

Reducing M from 5 to 1 improves the mean FID and KID of $\mathcal{D}_\beta(\beta{=}9)$, with gains in R1 and R2 and losses in R3. At $M = 1$, $\mathcal{D}_\beta(\beta{=}9)$ still improves FID over MFD in all three runs and KID in R1 and R2; R3's KID is slightly higher (0.641 versus 0.632).

Additive $\mathcal{D}_0$ responds oppositely: both means worsen at $M = 1$ and exceed those of $\mathcal{D}_\beta(\beta{=}9)$. Its KID is higher than $\mathcal{D}_\beta(\beta{=}9)$ in all three paired runs, reversing the ordering at $M = 5$. The preferred fitting budget therefore depends on the potential objective.

SANA. We evaluate $M \in \{1, 5, 10\}$ under the natural-magnitude SANA protocol of Appendix E.5, changing only the potential-update count. The $M = 5$ model and the 5000 held-out evaluation prompts are shared with Table 2.

Table 14: SANA potential-update count. Checkpoint 3000, fixed 5k prompts. M is the number of potential updates before each student update. The two panels use the same four models and images.

Method	FID ↓	Aesthetic ↑
MFD	27.048	6.964
FWCD, $M = 5$	24.755	**7.011**
FWCD, $M = 1$	24.675	6.949
FWCD, $M = 10$	**21.634**	6.857

Method	PickScore ↑	HPSv2 ↑	CLIPScore ↑	ImageReward ↑
MFD	0.1997	0.2344	29.079	0.429
FWCD, $M = 5$	**0.2007**	**0.2434**	**29.541**	**0.512**
FWCD, $M = 1$	0.1989	0.2361	29.078	0.505
FWCD, $M = 10$	0.1986	0.2402	29.226	0.468

FID and preference scores favor different budgets. $M = 10$ lowers FID by 12.61% relative to $M = 5$, or 20.02% relative to MFD, while all five preference scores peak at $M = 5$. $M = 1$ lowers FID by just 0.32% relative to $M = 5$ and reduces all five preference scores.

These comparisons match student updates, not potential updates or wall-clock time. ImageNet and SANA also differ in condition type, magnitude rule, batch construction, and reference source, so the results do not isolate task difficulty as a basis for choosing M.

I.4 Potential objective and correspondence: Shapley decomposition

The 2×2 comparison separates potential objective and correction correspondence under the common $0.9b$ base, 10% cap, and $M = 5$. Factor B switches from $\mathcal{D}_0$ to $\mathcal{D}_\beta(\beta{=}9)$, including the objective and its associated gradient scaling. Factor A switches from within-condition shuffling to the original query–correction correspondence. All four cells use the 2000-step ImageNet protocol for R1–R3 and the 3000-step SANA protocol for S1. ImageNet uses 50k samples per run; SANA uses the existing 5k prompts and teacher-image reference.

For each metric, let $v_{00}, v_{10}, v_{01}, v_{11}$ denote $\mathcal{D}_0$ shuffled, $\mathcal{D}_\beta(\beta{=}9)$ shuffled, $\mathcal{D}_0$ aligned, and $\mathcal{D}_\beta(\beta{=}9)$ aligned, respectively. Set v to the negative of FID, FD-DINOv2, or KID, and to the positive value of each preference score. The two-factor Shapley values are

$$\begin{aligned} \phi_B &= \tfrac{1}{2}[(v_{10} - v_{00}) + (v_{11} - v_{01})], \\ \phi_A &= \tfrac{1}{2}[(v_{01} - v_{00}) + (v_{11} - v_{10})], \qquad \phi_B + \phi_A = v_{11} - v_{00}. \end{aligned} \tag{40}$$

The decomposition attributes the change from $\mathcal{D}_0$ shuffled to $\mathcal{D}_\beta(\beta{=}9)$ aligned, rather than the full gain over MFD. Contributions are reported separately in each metric's original units, with positive values indicating improvement. ImageNet reports the mean and sample standard deviation of contributions computed within R1–R3; SANA reports S1.

The four-cell FID/KID and preference measurements appear in Tables 7 and 2; Table 15 adds FD-DINOv2, the Fréchet distance in DINOv2-large features (Oquab et al., 2024). SANA uses the final CLS token, sample covariance, and the same 5000 teacher images for all four cells. All contributions are computed from the original measurements before table rounding.

Table 15: Additional FD-DINOv2 values for the four-cell Shapley analysis. ImageNet uses 50k samples per replicate and real-reference statistics; SANA uses 5k samples and the common teacher-image reference. Lower is better.

	ImageNet			SANA
Configuration	R1	R2	R3	S1
$\mathcal{D}_0$ shuffled	103.1890	111.1840	**96.2276**	275.9402
$\mathcal{D}_0$ aligned	**95.3036**	**104.2010**	109.3810	256.5527
$\mathcal{D}_\beta(\beta{=}9)$ shuffled	110.0050	122.6180	109.0730	**225.8383**
$\mathcal{D}_\beta(\beta{=}9)$ aligned	107.0930	111.7210	100.2410	230.7268

Table 16: Shapley contributions in metric units. ImageNet: mean $\pm$ sample standard deviation over R1–R3. SANA: S1. Total is $v_{11} - v_{00}$. Calculations use the original recorded values; contributions and standard deviations are displayed to three significant figures.

Task	Metric	ϕ_B: objective	ϕ_A: correspondence	Total
ImageNet	FID	-0.144 ± 0.168	$+0.238 \pm 0.204$	$+0.0945 \pm 0.120$
	FD-DINOv2	-6.88 ± 4.35	$+4.06 \pm 5.67$	-2.82 ± 1.98
	KID $\times 1000$	-0.0230 ± 0.0536	$+0.0522 \pm 0.0338$	$+0.0292 \pm 0.0422$
SANA	FID	$+3.94$	$+1.14$	$+5.08$
	FD-DINOv2	$+38.0$	$+7.25$	$+45.2$
	Aesthetic	-0.00902	$+0.0548$	$+0.0457$
	PickScore	$+0.00105$	$+0.000468$	$+0.00152$
	HPSv2	$+0.00159$	$+0.00128$	$+0.00287$
	ImageReward	$+0.0755$	$+0.0268$	$+0.102$
	CLIPScore	$+0.119$	$+0.111$	$+0.230$

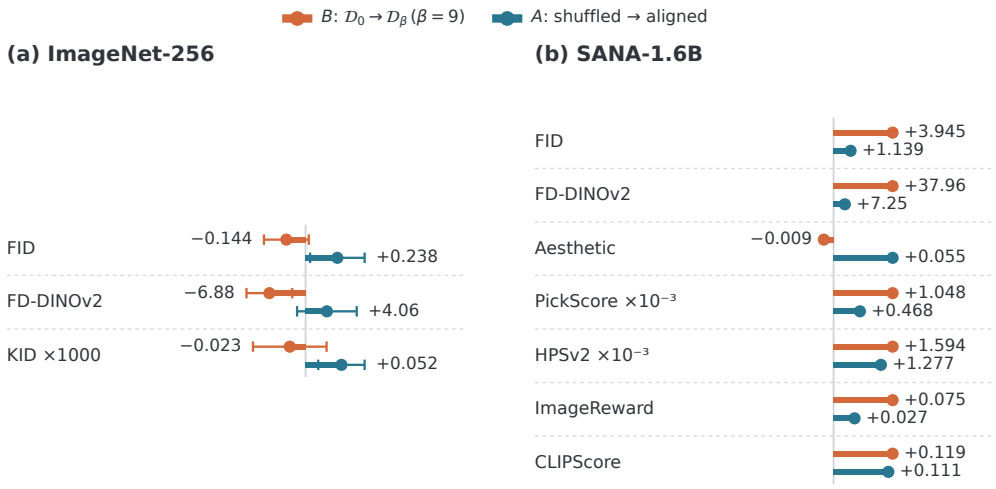

Figure 8: Shapley contributions of potential objective (B: $\mathcal{D}_0 \to \mathcal{D}_\beta(\beta{=}9)$, orange) and query–correction correspondence (A: shuffled to aligned, blue). Dots mark the estimates; ImageNet error bars show sample standard deviations across three training runs, while SANA uses one run. Each metric uses an independent scale, with positive values indicating improvement; PickScore and HPSv2 are displayed in units of 10^{-3}. The decomposition attributes the change from shuffled Additive $\mathcal{D}_0$ to aligned $\mathcal{D}_\beta(\beta{=}9)$, rather than the total gain over MFD.

The objective contribution differs by task: switching to $\mathcal{D}_\beta(\beta{=}9)$ has negative means on all three ImageNet metrics, but positive SANA contributions on six metrics and a slightly negative contribution for Aesthetic. Correspondence contributes positively to ImageNet FID and KID in R1–R3, to FD-DINOv2 in two runs, and to all seven SANA metrics. These averages coexist with conditional reversals: SANA alignment lowers Aesthetic, HPSv2, and ImageReward under $\mathcal{D}_0$, and raises FD-DINOv2 under $\mathcal{D}_\beta(\beta{=}9)$.

J Supplementary Controls

J.1 Multistep base model and four-step students

Table 17: Multistep base model versus four-step students on one common 10k evaluation stream with CFG 1. FWCD uses $\beta = 9$ on replicate R1. Lower is better.

Model	NFE	FID ↓	KID ×1000 ↓
SiT-XL/2 base, multistep	50	11.4507	6.36123
SiT-XL/2 base, naive four-step	4	94.1200	79.4456
MFD, checkpoint 2000	4	7.42414	0.647570
FWCD, checkpoint 2000	4	**7.16179**	**0.578493**

Naive four-step integration of the base model loses substantial quality, and distillation closes most of that gap. Under this sampler and CFG 1, the distilled students also have lower point estimates than the 50-step base model.

J.2 Empirical pairing displacements on two moons

The density constraint of Section 3 assigns no correspondence between samples. As a control, we form the correction directly from empirical pairing displacements. The setting uses the two-moons teacher, student initialization, 50:20 schedule, four-step sampling, and evaluation below, but trains no potential.

The two-dimensional source is an eight-component Gaussian mixture,

$$x_0 \sim \frac{1}{8}\sum_{j=1}^{8}\mathcal{N}(5v_j, 0.1I_2),$$

where v_j are the four coordinate-axis directions and four normalized diagonal directions. Conditions $c \in \{0,1\}$ identify the two moons. With $\vartheta \sim \mathrm{Unif}[0,\pi]$ and $\epsilon \sim \mathcal{N}(0, 0.04I_2)$, targets are

$$z^0(\vartheta) = (\cos\vartheta, \sin\vartheta), \qquad z^1(\vartheta) = (1-\cos\vartheta, 0.5-\sin\vartheta), \tag{41}$$

$$x_1^c = 3(z^c(\vartheta)+\epsilon) - \mathbf{1}. \tag{42}$$

Each repeated experiment trains its own conditional teacher of width 128 with three residual blocks for 20,000 straight FM steps at batch 256, and shares that teacher across methods in the repeat. Students and auxiliary models initialize from it and update all parameters at batch 256 on the 50:20 schedule. The four-step students receive 560 updates out of 2000 global steps. Evaluation uses 10,000 class-balanced samples; empirical W_2^2 uses the first 1000 samples; squared sliced Wasserstein distance uses 256 fixed projections; MMD uses five RBF bandwidths.

At each student update, a Hungarian assignment $\pi^\star$ within each class matches current student queries to independent reference queries, and the student direction is

$$g_i^{\mathrm{OT}} = b_i + x^Q_{\pi^\star(i)} - x_i^P.$$

Raw Random assigns reference samples at random within class and rescales the resulting displacements to match the classwise mean and RMS of the OT displacements. Both arms keep the natural displacement magnitude; the mean correction-to-base RMS ratios are 0.976 for Raw OT and 0.928 for Raw Random. Raw OT-4 and Raw Random-4 in Table 18 denote MFD plus the respective displacement. The MFD-4 row is the MFD arm trained in this campaign.

Table 18: Two-moons endpoints with empirical pairing displacements over ten training seeds. Mean $\pm$ between-seed standard deviation; lower is better. Bold marks the best four-step method.

Method	W_2^2	SWD^2	MMD^2	Teacher MSE
Teacher-50	0.1817 ± 0.0527	0.0392 ± 0.0203	0.00157 ± 0.00078	—
Naive-4	0.7743 ± 0.1174	0.2465 ± 0.0492	0.02013 ± 0.00320	$\mathbf{0.6108} \pm 0.0350$
MFD-4	2.6541 ± 1.6876	1.1103 ± 0.7911	0.04177 ± 0.02690	2.6612 ± 1.5758
Raw OT-4	$\mathbf{0.4990} \pm 0.1368$	$\mathbf{0.1374} \pm 0.0837$	$\mathbf{0.01206} \pm 0.00413$	0.8009 ± 0.2850
Raw Random-4	0.8576 ± 0.2646	0.2254 ± 0.0909	0.02655 ± 0.00815	1.4210 ± 0.6931

Raw OT is lower than MFD on all four metrics in 10/10 seeds, with a paired W_2^2 difference of -2.1551 and 95% interval $[-3.4469, -0.8632]$. Against Raw Random, W_2^2, MMD^2, and teacher-endpoint MSE are lower in 8/10, 10/10, and 9/10 seeds with intervals below zero; SWD^2 is lower in 9/10 seeds with interval $[-0.1872, 0.0112]$.

For the direction comparison in Figure 9, the reference field is a 64×64 weighted Poisson solve on $[-8,8]^2$: student and reference histograms are Gaussian-smoothed with standard deviation one grid cell, the student density receives an additive floor of 10^{-4} times its mean, and finite differences of the solved potential are interpolated to query positions. At $t = 0.2, 0.5, 0.8$, the mean cosine between Raw OT displacements and this field is $0.692, 0.631, 0.753$, versus $0.399, 0.253, 0.230$ for Raw Random. In this two-dimensional setting, empirical pairing displacements carry directional information and improve the endpoint distribution.

J.3 Empirical pairing displacements on ImageNet

ImageNet uses the query construction of Appendix E.4. Each global batch holds two classes with 16 queries each. For each class we form the 16×16 cost $C_{ij} = \|x_i^P - x_j^Q\|^2/d$, solve the Hungarian assignment $\pi_c^\star$, and set $d_i^{\mathrm{OT}} = x^Q_{\pi_c^\star(i)} - x_i^P$. With $\mathcal{R}(\cdot)$ the global-batch RMS, three magnitudes are compared:

$$\delta_i^{\mathrm{Raw}} = d_i^{\mathrm{OT}}, \qquad \delta_i^{5\%} = 0.05\,\frac{\mathcal{R}(b)}{\mathcal{R}(d^{\mathrm{OT}})}\,d_i^{\mathrm{OT}}, \qquad \delta_i^{10\%} = 0.1\min\Big\{1, \frac{\mathcal{R}(b)}{\mathcal{R}(d^{\mathrm{OT}})}\Big\}d_i^{\mathrm{OT}}.$$

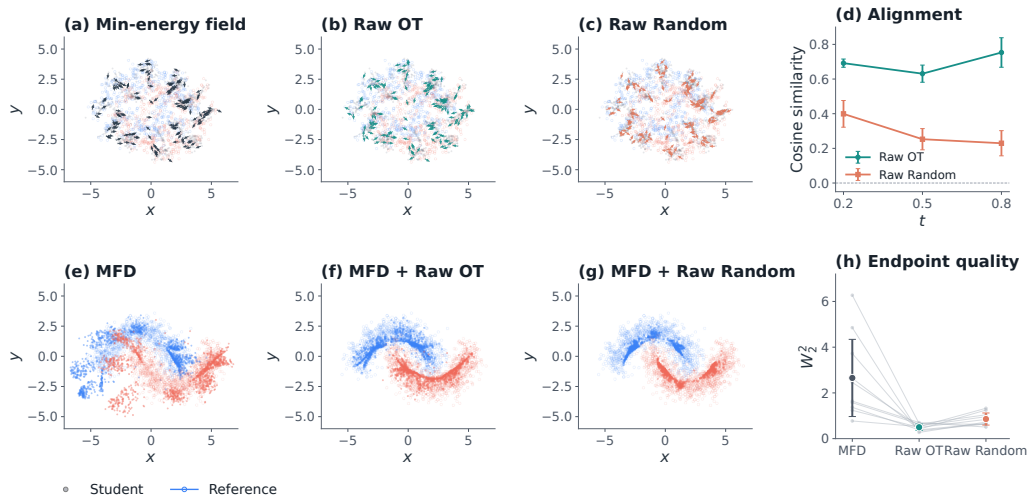

Figure 9: Two-moons pairing control. Top: finite-grid reference field at $t = 0.5$, Raw OT displacement, Raw Random displacement matched in classwise mean and RMS, and ten-seed direction cosines. Bottom: four-step endpoints for one fixed seed under MFD, Raw OT, and Raw Random, with paired W_2^2. Blue/red distinguish the two conditions. Arrows are normalized for display only.

The student direction is $g_i = b_i + \delta_i$, with displacement and b under stop-gradient. Random controls permute reference samples within class and match the classwise mean and centered RMS of the OT displacement before applying the same magnitude rule. Replicate R6 is a single-run control (Table 5); checkpoint 2000, 50k images.

Table 19: ImageNet pairing displacements at three magnitudes, 50k endpoint evaluation. Lower is better.

Method	FID ↓	KID ×1000 ↓
MFD	**4.117**	**0.685**
Raw OT	243.028	212.569
Raw Random	256.372	236.099
Exact-5% OT	4.125	0.707
Exact-5% Random	4.202	0.942
10% cap OT	4.236	0.801

The Raw OT displacement has mean RMS about 3.95 times the base direction, with minimum about 1.51 and maximum above 152; adding it directly degrades both metrics severely. Raw OT beats the mean- and RMS-matched Raw Random on both metrics, so the pairing carries some directional information, but it remains far worse than MFD. The 5% relative magnitude avoids the degradation and is briefly better than MFD at the 10k checkpoint-1000 evaluation (8.819 versus 9.044), but at the pre-registered checkpoint-2000 50k endpoint it does not improve FID or KID over MFD, and it has no consistent advantage over the magnitude-matched Random control. The 10% cap also does not improve the two endpoint metrics.

The two-moons result does not transfer to this real-image setting. The direction and magnitude of a pairing displacement are hard to separate in high dimension, and the raw empirical pairing is not a usable correction. FWCD instead learns a potential gradient from the two marginals and prescribes the magnitude rule separately.

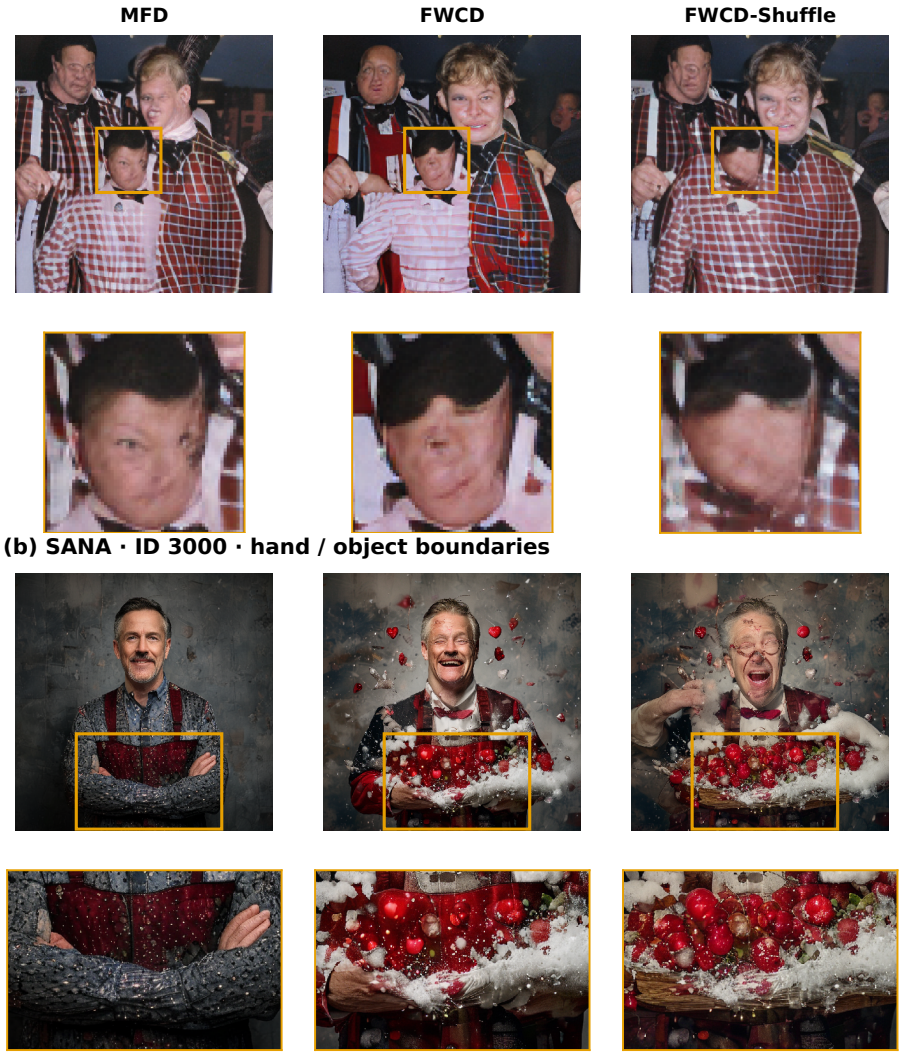

Figure 10: Local-structure examples from (a) ImageNet R2, index 028571, and (b) SANA natural-magnitude arms, prompt ID 3000. Columns compare MFD, FWCD, and FWCD-Shuffle under shared conditions and initial noise. Full images and orange-box enlargements use identical crops and nearest-neighbor display across methods. These selected examples do not estimate failure rates.

J.4 QUALITATIVE FAILURE CASES

FWCD leaves the small foreground face unresolved in ImageNet and the hand–object boundary ambiguous in SANA; the shuffled SANA output also has conspicuous facial distortion. These local defects persist despite aggregate quality gains. Marginal supervision provides no direct constraint on anatomy or object boundaries, and these selected examples do not establish a query-mismatch cause.

K COCO REAL-REFERENCE AND PREFERENCE EVALUATION

We evaluate the existing MFD and capped $\mathcal{D}_\beta(\beta{=}9)$ FWCD SANA models on COCO2014 validation (Lin et al., 2014): S1, checkpoint 3000, 1024-pixel generation, four steps and CFG 1, with no additional training.

Sort the 40,504 validation images by image ID and retain the caption with the smallest caption ID for each image. Normalize whitespace, apply a CPU `torch.randperm` initialized with s_{S1} (Table 5), and take the first 30,000 entries. The models share this ordering and use image seed $s_{\mathrm{S1}} + i$ for sample index i. The selected captions have no exact match, after whitespace normalization and case folding, to either the 3000 training texts or the full 479,410-text pool. This check does not resolve semantic overlap or overlap with the base model's pretraining data.

30k distribution and text alignment. FID compares 30,000 generated images with all 40,504 real validation images using clean-fid 0.1.35, clean mode and Inception-v3.[5] Generated RGB images are first downsampled to 256×256 by PIL bicubic interpolation; real JPEGs receive no pre-resize. Both then undergo standard clean-fid preprocessing. CLIPScore uses the original generated images and CLIP ViT-B/32 as in Appendix E. These real-reference scores are separate from teacher-reference SANA, published COCO10k/512-pixel FID, and OpenCLIP-G scores.

Table 20: COCO2014 evaluation of existing SANA checkpoints. FID and CLIPScore both use 30k generated images; FID has 40,504 real reference images. S1 for each method.

Method	Images	FID ↓	CLIPScore ↑
MFD	30,000	27.470	30.497
FWCD ($\beta = 9$, cap)	30,000	**20.729**	**30.866**

Shared 5k preference subset. All five scores use indices 0–4999 of the same randomized list, with shared prompts and noise. For computational cost, the preference subset was reduced from 30k to 5k after the 30k FID and CLIPScore results were available; no indices were selected by quality or score. Aesthetic and CLIPScore reuse saved per-image values. PickScore, HPSv2.1 and ImageReward use the same subset and the evaluator settings of Appendix E, batch size eight and FP32.

Table 21: COCO preference scores on the common 5k subset. All columns use the same 5000 samples; all are higher-is-better. HPSv2 uses v2.1 weights.

Method	Aesthetic	PickScore	HPSv2	ImageReward	CLIPScore
MFD	6.582	0.2242	0.2678	0.707	30.471
FWCD ($\beta = 9$, cap)	**6.693**	**0.2257**	**0.2802**	**0.821**	**30.839**

Compared with MFD, FWCD reduces FID by 6.742 (24.54%), raises 30k CLIPScore by 0.369, and improves all five 5k scores. These are single-training-seed point comparisons.

[5] `https://github.com/GaParmar/clean-fid`