THREE-BODY SCATTERING FOR GENERATIVE MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

An energy-distance field provides direct sample-level supervision for one-step generation, but sparse source interactions are noisy. We propose Three-Body Scattering Modeling (TBSM), which learns a fake-to-real correction field through an online tracker that adapts to the evolving generated distribution. Each generated projectile is attracted toward one real source and repelled from one independently generated source. At the exact energy-distance endpoint, the conditional expectation of the interaction equals the 2-Wasserstein gradient-flow velocity of $\frac{1}{2}D_E^2(P_{\boldsymbol{\theta}}, Q)$. The tracker approximates this expectation; frozen-target regression transfers its corrections into the generator, which is used alone at inference. Our analysis characterizes when tracking reduces supervision error, with stationarity and convergence results under explicit assumptions. Each event uses one real reference for its condition, yielding $O(B)$ sample interactions for a batch of B projectiles. Using frozen image features, TBSM trains one-step generators on ImageNet-256, achieving FID $= 2.23$ with pixel-space PixelDiT-XL and FID $= 1.63$ with latent-space DiT-XL at NFE $= 1$. A design map relates this learning loop to diffusion-related supervision, Drift-like dynamics, and GAN-like objectives.



Figure 1: **Samples generated in one step by TBSM-trained models.** From left to right: MNIST, Fashion-MNIST, and CIFAR-10 (top to bottom); ***pixel-space*** ImageNet-1K samples; and text-to-image samples generated at **NFE** $= 1$ **by a Qwen-Image-20B (Wu et al., 2025) model fine-tuned using only our TBSM (see App. C).**

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

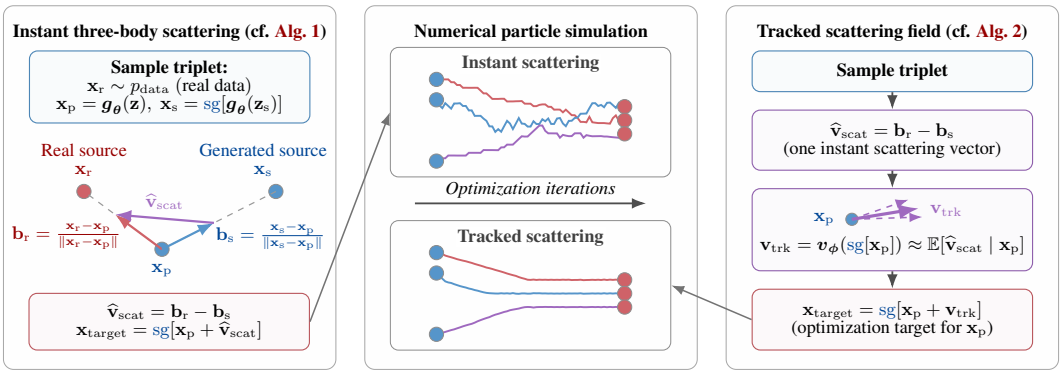


Figure 2: **Overview of TBSM.** Left: a real–generated–generated event yields a noisy scattering vector. Center: a particle simulation illustrates noisy instant updates and smoother tracked trajectories. Right: at the fully tracked energy-distance endpoint, an online tracker estimates the conditional expectation of the instant scattering vector at each projectile and supervises the generator. Center-panel red, blue, and purple curves show generated-particle paths from blue initial points; red dots mark targets; sg[·] denotes stop gradient.

1 Introduction

Generative models rely on adversarial comparison (Goodfellow et al., 2014; Brock et al., 2018), path-indexed diffusion, score, or flow supervision (Ho et al., 2020; Song et al., 2020b; Lipman et al., 2022), or autoregressive prediction (van den Oord et al., 2016; Tian et al., 2024; Sun et al., 2024). Direct one-step distribution-matching methods usually estimate supervision from minibatch fields, statistics, or variational objectives (Deng et al., 2026; Yang et al., 2026; Caucheteux et al., 2026). Classical linear-time distance-kernel MMD already admits constant-size stochastic updates (Gretton et al., 2012); their source-sampling noise motivates learning a field that can provide more stable generator supervision.

> **Problem 1** . *Can an online tracker turn sparse, noisy sample interactions into a stable fake-to-real correction field for high-dimensional one-step generation?*

We propose Three-Body Scattering Modeling (TBSM) to learn a *fake-to-real correction field* from these noisy interactions. Across training iterations, an online tracker fits their conditional mean while adapting to the evolving generated distribution. Frozen-target regression transfers the learned corrections into a one-step generator, which is used alone at inference (Fig. 2).

Each training event treats a generated sample as a projectile attracted to one real source and repelled from an independent generated source. Averaging this interaction recovers the 2-Wasserstein descent velocity of the proper energy-distance objective $\frac{1}{2}D_E^2(P_{\boldsymbol{\theta}}, Q)$ (Székely & Rizzo, 2013; Ambrosio et al., 2008); one triplet per projectile gives $O(B)$ detached-source targets for B projectiles. We call this supervision sample-level because each condition-specific loss requires only one paired real observation, naturally fitting text–image datasets with one image per caption rather than same-condition reference sets.

The tracker turns sparse observations into learned supervision through a detached-target interface (Deng et al., 2026; Yang et al., 2026). At the exact endpoint, oracle tracking preserves the population field and removes source-sampling variance; finite-error decompositions quantify when a learned tracker improves this supervision.

The exact energy-distance field anchors the theory; experiments retain the regression loop in frozen representation spaces with smoothed bearings and sometimes reweighted intra-source interaction. The TBSM objective requires no path-model teacher queries, prescribed denoising target, adversarial discriminator, or ordered factorization. Training and inference use no classifier-free guidance or guidance-scale sweep; App. D contrasts pipelines that bake guidance into a training field or distillation target. Primary ImageNet runs use multi-step checkpoints only for initialization (App. B). Tab. 5 in the appendix compares paradigms by sampling form, scaling evidence, and supervision granularity. At

NFE $= 1$, PixelDiT-XL (pixel) and DiT-XL/2 (latent) reach FID $= 2.23$ and 1.63 on ImageNet-256; DiT-XL/4 reaches 1.92 on ImageNet-512. Fig. 1 adds small-scale and text-to-image samples.

Our analysis identifies two roles for tracking. At the energy-distance endpoint, interactions remain noisy even after matching a continuous data distribution; learning their conditional expectation filters this noise when tracker error is sufficiently small. At the full-corridor endpoint, the smoothed normalized field vanishes exactly when the endpoint laws match. These results explain the learned supervision; convergence requires additional flow or parameter-space assumptions (App. F.3). Our contributions are:

(a) **Learned fake-to-real supervision.** We train a tracker on sparse interactions to guide a one-step generator, characterize when it filters persistent sampling noise, and prove that its smoothed full-corridor population field identifies distributional agreement.
(b) **High-dimensional tracked generation.** We apply this loop in frozen representation spaces for pixel- and latent-output generators. Matched-update comparisons favor tracking, and ImageNet benchmarks demonstrate high-quality one-step generation.
(c) **An interpretive design map.** We connect TBSM to Drift-like dynamics, a path-regularized GAN-like displacement analogue, and diffusion-related supervision via input and target changes.

2 Methodology

We develop an online tracker to learn generator corrections. We first establish its stochastic supervision and frozen-target interface (Alg. 1), then learn the field and analyze tracking error (Alg. 2).

2.1 Distributional Objective and Stochastic Optimization

Population objective. Let $Q(\cdot \mid \mathbf{c})$ be the real probability law on $\mathbb{R}^d$ and let $\boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{c})$ be a conditional generator with $\mathbf{z} \sim p(\mathbf{z})$ and optional condition $\mathbf{c}$. For each condition, the generator induces the probability law $P_{\boldsymbol{\theta}}(\cdot \mid \mathbf{c}) = (\boldsymbol{g}_{\boldsymbol{\theta}}(\cdot, \mathbf{c}))_{\#}p(\mathbf{z})$; in the unconditional case we write these laws as $P_{\boldsymbol{\theta}}$ and Q. We use $\mathbf{x}$ for a generic spatial argument of a potential or vector field and reserve $\mathbf{x}_{\mathrm{p}}$ for a sampled generated projectile in a stochastic three-body event. The learning goal is condition-wise distribution matching:

$$\mathcal{F}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{c}\sim p_{\mathrm{data}}(\mathbf{c})}\left[\frac{1}{2}D_E^2\left(P_{\boldsymbol{\theta}}(\cdot \mid \mathbf{c}), Q(\cdot \mid \mathbf{c})\right)\right]. \tag{1}$$

The unconditional case is recovered by dropping $\mathbf{c}$. For finite paired datasets with one observation per condition, $Q(\cdot \mid \mathbf{c})$ is a population object rather than the literal empirical distribution for that condition. We use the squared energy distance as the population discrepancy:

$$\begin{aligned} D_E^2(P_{\boldsymbol{\theta}}, Q) = 2\mathbb{E}_{\mathbf{x}_{\mathrm{p}}\sim P_{\boldsymbol{\theta}}, \mathbf{x}_{\mathrm{r}}\sim Q}\left[\|\mathbf{x}_{\mathrm{p}} - \mathbf{x}_{\mathrm{r}}\|\right] \\ - \mathbb{E}_{\mathbf{x}_{\mathrm{p}}, \mathbf{x}_{\mathrm{s}}\sim P_{\boldsymbol{\theta}}}\left[\|\mathbf{x}_{\mathrm{p}} - \mathbf{x}_{\mathrm{s}}\|\right] - \mathbb{E}_{\mathbf{x}_{\mathrm{r}}, \mathbf{x}'_{\mathrm{r}}\sim Q}\left[\|\mathbf{x}_{\mathrm{r}} - \mathbf{x}'_{\mathrm{r}}\|\right]. \end{aligned} \tag{2}$$

Euclidean space is of strong negative type, so $D_E^2(P_{\boldsymbol{\theta}}, Q) \geq 0$ and equality holds exactly when $P_{\boldsymbol{\theta}} = Q$ under finite first moments. Equivalently, $\frac{1}{2}D_E^2$ is a squared distance-kernel MMD; App. D.5 states its precise relation to the stochastic update below.

Population vector field. For a fixed condition, suppress $\mathbf{c}$ and write $\mathcal{F}(P_{\boldsymbol{\theta}}) = \frac{1}{2}D_E^2(P_{\boldsymbol{\theta}}, Q)$. The 2-Wasserstein negative-gradient flow of a functional $\mathcal{F}(P)$ has velocity $\mathbf{v}(\mathbf{x}) = -\nabla_{\mathbf{x}}\frac{\delta \mathcal{F}}{\delta P}(\mathbf{x})$. For the energy distance, differentiating its first variation condition-wise gives the generated-particle scattering vector field

$$\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{c}) = \mathbb{E}_{\mathbf{x}_{\mathrm{r}}\sim Q(\cdot\mid\mathbf{c})}\left[\frac{\mathbf{x}_{\mathrm{r}} - \mathbf{x}}{\|\mathbf{x}_{\mathrm{r}} - \mathbf{x}\|}\right] - \mathbb{E}_{\mathbf{x}_{\mathrm{s}}\sim P_{\boldsymbol{\theta}}(\cdot\mid\mathbf{c})}\left[\frac{\mathbf{x}_{\mathrm{s}} - \mathbf{x}}{\|\mathbf{x}_{\mathrm{s}} - \mathbf{x}\|}\right]. \tag{3}$$

These inter- and intra-source expected bearings have unit-norm integrands away from coincidence. The model–model term contributes through both arguments, yielding the unit repulsion coefficient; App. F.1 gives the first variation and generator-Jacobian pullback. For estimator identities we set each bearing to zero at exact coincidence; code uses denominator smoothing, which is the exact gradient of a differentiable radial potential and converges pointwise to the bearing field as detailed in App. F.5.1. The ideal flow follows $\mathrm{d}\mathbf{x}/\mathrm{d}t = \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{c})$; Thm. 1 below establishes the corresponding generator update.

Algorithm 1 Instant Three-Body Scattering Regression

1: Initialize generator parameters $\boldsymbol{\theta}$.
2: **repeat**
3: Sample real source-condition pairs $(\mathbf{x}_\mathrm{r}, \mathbf{c})$ from dataset.
4: Independently sample $\mathbf{z}, \mathbf{z}_\mathrm{s} \sim p(\mathbf{z})$, and generate $\mathbf{x}_\mathrm{p} = \boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{c})$ and $\mathbf{x}_\mathrm{s} = \mathrm{sg}[\boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}_\mathrm{s}, \mathbf{c})]$.
5: Compute bearings $\mathbf{b}_\mathrm{r} = (\mathbf{x}_\mathrm{r} - \mathbf{x}_\mathrm{p}) / \|\mathbf{x}_\mathrm{r} - \mathbf{x}_\mathrm{p}\|$ and $\mathbf{b}_\mathrm{s} = (\mathbf{x}_\mathrm{s} - \mathbf{x}_\mathrm{p}) / \|\mathbf{x}_\mathrm{s} - \mathbf{x}_\mathrm{p}\|$.
6: Update $\boldsymbol{\theta}$ by descending $\|\mathbf{x}_\mathrm{p} - \mathrm{sg}[\mathbf{x}_\mathrm{p} + \mathbf{b}_\mathrm{r} - \mathbf{b}_\mathrm{s}]\|_2^2$.
7: **until** convergence

Stochastic estimator. A full empirical evaluation of (2) uses pairwise distances, whereas classical random-pair MMD estimators are linear in batch size (Gretton et al., 2012). We use one side of this stochastic gradient as a projectile-level field with one differentiable projectile, one real source, one detached generated source, and an explicit target that an auxiliary field can track; its $O(B)$ scaling refers to sample interactions, not per-sample network cost. The exact relation to the symmetrized estimator is given in App. D.5.

2.2 Three-Body Scattering Vector Field

Triplet estimator. We estimate (3) with a constant-size signed interaction; App. F.4 makes precise why a real-only estimator cannot recover its model-dependent term. For a condition $\mathbf{c}$ and projectile $\mathbf{x}_\mathrm{p} \sim P_{\boldsymbol{\theta}}(\cdot \mid \mathbf{c})$, sample one real source $\mathbf{x}_\mathrm{r} \sim Q(\cdot \mid \mathbf{c})$ and one independently generated source $\mathbf{x}_\mathrm{s} \sim P_{\boldsymbol{\theta}}(\cdot \mid \mathbf{c})$. The corresponding sample-level scattering estimator is

$$\widehat{\mathbf{v}}_\mathrm{scat}(\mathbf{x}_\mathrm{p}; \mathbf{x}_\mathrm{r}, \mathbf{x}_\mathrm{s}) = \mathbf{b}_\mathrm{r} - \mathbf{b}_\mathrm{s}, \qquad \mathbf{b}_\mathrm{r} = \underbrace{\frac{\mathbf{x}_\mathrm{r} - \mathbf{x}_\mathrm{p}}{\|\mathbf{x}_\mathrm{r} - \mathbf{x}_\mathrm{p}\|}}_{\text{inter-source bearing}}, \qquad \mathbf{b}_\mathrm{s} = \underbrace{\frac{\mathbf{x}_\mathrm{s} - \mathbf{x}_\mathrm{p}}{\|\mathbf{x}_\mathrm{s} - \mathbf{x}_\mathrm{p}\|}}_{\text{intra-source bearing}}, \tag{4}$$

Averaging over the two sources recovers the population scattering vector field:

$$\mathbb{E}_{\mathbf{x}_\mathrm{r} \sim Q(\cdot|\mathbf{c}),\, \mathbf{x}_\mathrm{s} \sim P_{\boldsymbol{\theta}}(\cdot|\mathbf{c})} \left[\widehat{\mathbf{v}}_\mathrm{scat}(\mathbf{x}_\mathrm{p}; \mathbf{x}_\mathrm{r}, \mathbf{x}_\mathrm{s})\right] = \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}_\mathrm{p}, \mathbf{c}). \tag{5}$$

Away from coincidence, the same vector is the negative projectile gradient of the single-event energy $\|\mathbf{x}_\mathrm{p} - \mathbf{x}_\mathrm{r}\| - \|\mathbf{x}_\mathrm{p} - \mathbf{x}_\mathrm{s}\|$, whose source expectation is the first variation in (15) evaluated at the projectile.

Scattering interpretation. Here, three-body denotes a projectile's signed interaction with real and model sources, not Newtonian dynamics; Thm. 1 formalizes its objective-to-update identity.

2.3 Scattering Target Regression

Frozen target. For each loss event, we sample a projectile, real source, and independently generated source, displace the projectile by the sampled inter-minus-intra vector, and regress the generator output toward this detached target. The one-step target is

$$\mathbf{x}_\mathrm{target} = \mathrm{sg}\left[\mathbf{x}_\mathrm{p} + \widehat{\mathbf{v}}_\mathrm{scat}(\mathrm{sg}[\mathbf{x}_\mathrm{p}]; \mathbf{x}_\mathrm{r}, \mathbf{x}_\mathrm{s})\right], \qquad \mathbf{x}_\mathrm{p} = \boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{c}), \quad \mathbf{x}_\mathrm{s} = \mathrm{sg}[\boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}_\mathrm{s}, \mathbf{c})]. \tag{6}$$

The generator objective averages this target regression over triplets:

$$\mathcal{L}_\mathrm{gen}(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}_{(\mathbf{x}_\mathrm{r}, \mathbf{c}) \sim p_\mathrm{data}(\mathbf{x}, \mathbf{c}),\, \mathbf{z}, \mathbf{z}_\mathrm{s} \sim p(\mathbf{z})} \left[\left\|\boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{c}) - \mathbf{x}_\mathrm{target}\right\|_2^2\right]. \tag{7}$$

In Alg. 1, detaching the entire target keeps the sampled correction fixed during the generator update. Gradients then pass only through the projectile prediction, pulling this correction back through the generator Jacobian. Differentiating through the target would change that update by introducing derivatives of the target's projectile and field. This separation also lets a learned field supply the correction through the same regression interface.

Local gradient identity. Fix $\mathbf{c}$ and current parameters $\bar{\boldsymbol{\theta}}$. Let $\mathcal{F}_\mathbf{c} = \frac{1}{2} D_E^2(P_{\boldsymbol{\theta}}(\cdot \mid \mathbf{c}), Q(\cdot \mid \mathbf{c}))$ and let $\mathcal{L}_{\mathrm{gen},\mathbf{c}}(\boldsymbol{\theta}; \bar{\boldsymbol{\theta}})$ denote (7) with the condition fixed and targets frozen at $\bar{\boldsymbol{\theta}}$. Assume independent

projectile/source draws, exact bearings, finite first moments, zero collision probability, almost-sure generator differentiability, and valid differentiation under the expectation; App. F.1 gives the full notation and proof.

Theorem 1 (Energy-distance three-body local gradient equivalence) . *Under the conditions above, exact-bearing frozen-target regression obeys the local identity*

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{gen},\mathbf{c}}(\boldsymbol{\theta}; \bar{\boldsymbol{\theta}})\big|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}} = \nabla_{\boldsymbol{\theta}} \mathcal{F}_{\mathbf{c}}(\bar{\boldsymbol{\theta}}). \tag{8}$$

This is a local gradient identity; the regression loss value does not estimate the energy distance. One joint real–condition draw gives an unbiased gradient for (1), since the real–real term is $\boldsymbol{\theta}$-independent. Alg. 1 absorbs the conventional $1/2$ into its learning rate. Proof, exact variance, and flow dissipation are in App. F.1, App. F.4, and App. F.3.1.

2.4 Tracked Scattering Targets

Why tracking helps. A *scattering tracker* learns the conditional expectation of sparse interactions. Fix a condition and draw the projectile and both sources independently from the same non-atomic law $P = Q$. For exact bearings, $Y = \widehat{\mathbf{v}}_{\mathrm{scat}}$ satisfies

$$\mathbb{E}\left[Y \mid \mathbf{x}_{\mathrm{p}}\right] = 0, \qquad 1 \le V_Q := \mathbb{E}\left[\|Y\|_2^2\right] \le \frac{4}{3}. \tag{9}$$

The bound averages over projectiles (Prop. 2). Thus the oracle outputs zero correction at agreement while its minimum raw regression loss remains V_Q: a positive loss need not indicate an inaccurate field. In a Gaussian location model, the local signal-to-noise slope scales as $d^{-1/2}$ (App. F.2.4); bounded interaction noise can therefore obscure small corrections in high dimension.

Exact and weighted targets. Write $\widehat{\mathbf{v}}_\lambda = \mathbf{b}_{\mathrm{r}} - \lambda \mathbf{b}_{\mathrm{s}}$, where the coupled design parameter $\lambda \in [0, 1]$ controls both the intra-source coefficient and the tracker-query range defined below. At $\lambda = 1$, its projectile-conditioned expectation is the energy-distance field (3). For $\lambda < 1$, this source expectation is the negative-gradient field of a weighted functional $\mathcal{F}_\lambda$ that favors lower generated-sample dispersion and generally shifts the optimum away from Q (App. F.5.2).

Tracker query and mixing. The tracker is queried at $\tilde{\mathbf{x}} = \alpha \mathbf{x}_{\mathrm{r}} + (1 - \alpha)\,\mathrm{sg}[\mathbf{x}_{\mathrm{p}}]$, with $\alpha \sim \mathcal{U}(0, 1 - \lambda)$ and point mass at $\alpha = 0$ when $\lambda = 1$. The tracker observes the query and condition, not α separately. The tracker is trained by the regression objective $\mathcal{L}_{\mathrm{trk}}(\boldsymbol{\phi}) = \mathbb{E}\left[\|\boldsymbol{v}_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}, \mathbf{c}) - \mathrm{sg}[\widehat{\mathbf{v}}_\lambda]\|_2^2\right]$. The tracked-supervision weight $\rho \in [0, 1]$ then mixes the instantaneous and tracked vectors as $\widehat{\mathbf{v}}_{\mathrm{mix}} = (1 - \rho)\widehat{\mathbf{v}}_\lambda + \rho \boldsymbol{v}_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}, \mathbf{c})$. The generator target is still (6), with $\widehat{\mathbf{v}}_{\mathrm{scat}}$ replaced by $\widehat{\mathbf{v}}_{\mathrm{mix}}$. For a fixed current generator and square-integrable tracker, at $\lambda = 1$ define the conditional source variance $s_{\boldsymbol{\theta}}(\mathbf{x}_{\mathrm{p}}, \mathbf{c}) = \mathrm{tr}\,\mathrm{Var}(\widehat{\mathbf{v}}_{\mathrm{scat}} \mid \mathbf{x}_{\mathrm{p}}, \mathbf{c})$ and tracker error $e_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{p}}, \mathbf{c}) = \|\boldsymbol{v}_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{p}}, \mathbf{c}) - \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}_{\mathrm{p}}, \mathbf{c})\|_2^2$.

Corollary 1 (Finite-tracker error at the energy-distance endpoint) . *At $\lambda = 1$, the mixed scattering vector satisfies the exact decomposition*

$$\mathbb{E}\left[\|\widehat{\mathbf{v}}_{\mathrm{mix}} - \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}_{\mathrm{p}}, \mathbf{c})\|_2^2 \mid \mathbf{x}_{\mathrm{p}}, \mathbf{c}\right] = (1 - \rho)^2 s_{\boldsymbol{\theta}}(\mathbf{x}_{\mathrm{p}}, \mathbf{c}) + \rho^2 e_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{p}}, \mathbf{c}). \tag{10}$$

Full tracking improves the conditional field MSE exactly when $e_{\boldsymbol{\phi}} < s_{\boldsymbol{\theta}}$; the pointwise optimal mixture is $\rho^\star = s_{\boldsymbol{\theta}}/(s_{\boldsymbol{\theta}} + e_{\boldsymbol{\phi}})$ when the denominator is positive. Together with (9), this explains how learning can filter noise that persists even at agreement. Thm. 2 gives the full projection identity, and Cor. 2 quantifies error after generator-Jacobian pullback.

What the corridor field identifies. For $\lambda < 1$, conditioning on the corridor generally changes the expected correction because the query does not determine the projectile. At $\lambda = 0$, this yields a distributional criterion even with normalized targets.

Proposition 1 (Identifiability of the full-corridor bearing field) . *Fix a condition and any Borel laws P, Q on $\mathbb{R}^d$. Draw $X \sim P$, $R \sim Q$, and $A \sim \mathcal{U}[0, 1]$ independently, and set*

Algorithm 2 Full TBSM: Tracked Three-Body Scattering Regression

Initialize generator parameters $\boldsymbol{\theta}$ and tracker parameters $\boldsymbol{\phi}$.
Choose tracked-supervision weight $\rho \in [0, 1]$ and intra-source weight $\lambda \in [0, 1]$.
repeat
  Sample real source-condition pairs $(\mathbf{x}_\mathrm{r}, \mathbf{c})$ from dataset.
  Sample noise $\mathbf{z}, \mathbf{z}_\mathrm{s} \sim p(\mathbf{z})$, and generate $\mathbf{x}_\mathrm{p} = \boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{c})$ and $\mathbf{x}_\mathrm{s} = \mathrm{sg}[\boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}_\mathrm{s}, \mathbf{c})]$.
  Compute bearings $\mathbf{b}_\mathrm{r} = (\mathbf{x}_\mathrm{r} - \mathbf{x}_\mathrm{p})/\left\|\mathbf{x}_\mathrm{r} - \mathbf{x}_\mathrm{p}\right\|$ and $\mathbf{b}_\mathrm{s} = (\mathbf{x}_\mathrm{s} - \mathbf{x}_\mathrm{p})/\left\|\mathbf{x}_\mathrm{s} - \mathbf{x}_\mathrm{p}\right\|$.
  Form λ-weighted scattering vector $\widehat{\mathbf{v}}_\lambda = \mathbf{b}_\mathrm{r} - \lambda \mathbf{b}_\mathrm{s}$.
  Sample $\alpha \sim \mathcal{U}(0, 1 - \lambda)$ and set $\tilde{\mathbf{x}} = \alpha \mathbf{x}_\mathrm{r} + (1 - \alpha)\,\mathrm{sg}[\mathbf{x}_\mathrm{p}]$.
  Query tracked vector $\mathbf{v}_\mathrm{trk} = \boldsymbol{v}_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}, \mathbf{c})$.
  Form mixed scattering vector $\widehat{\mathbf{v}}_\mathrm{mix} = (1 - \rho)\widehat{\mathbf{v}}_\lambda + \rho \mathbf{v}_\mathrm{trk}$.
  Update $\boldsymbol{\theta}$ by descending $\left\|\mathbf{x}_\mathrm{p} - \mathrm{sg}[\mathbf{x}_\mathrm{p} + \widehat{\mathbf{v}}_\mathrm{mix}]\right\|_2^2$.
  Update $\boldsymbol{\phi}$ by descending $\left\|\mathbf{v}_\mathrm{trk} - \mathrm{sg}[\widehat{\mathbf{v}}_\lambda]\right\|_2^2$.
until convergence

$Z = (1 - A)X + AR$. For $\varepsilon > 0$, let $m_\varepsilon(Z) = \mathbb{E}\left[(R - X)/(\|R - X\|_2 + \varepsilon) \mid Z\right]$, without observing A separately. Then

$$\mathbb{E}\left[\|m_\varepsilon(Z)\|_2^2\right] = 0 \quad \Longleftrightarrow \quad P = Q. \tag{11}$$

At agreement, exchanging (X, R, A) with $(R, X, 1 - A)$ preserves the query and reverses the target. Instantaneous attraction can contract even matched laws; corridor conditioning cancels that correction. The converse shows that this cancellation cannot hide an endpoint mismatch, even after each displacement is normalized. Thus the query determines what distributional correction the tracker learns. This criterion concerns the full-corridor population field; it does not imply descent after generator pullback or extend to partial corridors (App. F.2.3).

Full tracked algorithm. Alg. 2 retains the gray sampling and bearing steps of Alg. 1 and adds source reweighting and tracked supervision. At $\rho = 0$, tracking may be omitted; $\lambda = 1$ then recovers Alg. 1. Experiments use this update structure with the representation interfaces in Sec. 2.5.

From a learned transfer field to one-step generation. Alg. 3 (App. E.1) makes the mechanism behind tracked supervision explicit. At $(\rho, \lambda) = (1, 0)$, it replaces the unit bearing with the displacement $\Delta = \mathbf{x}_\mathrm{r} - \mathbf{x}_\mathrm{p}$ and regresses it at fake-to-real corridor queries $\tilde{\mathbf{x}}$. The tracker thus learns $\mathbb{E}\left[\Delta \mid \tilde{\mathbf{x}}, \mathbf{c}\right]$, an average of sampled fake-to-real displacements. Regression toward $\mathrm{sg}[\mathbf{x}_\mathrm{p} + \boldsymbol{v}_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}, \mathbf{c})]$ transfers these corrections into the generator, while the tracker adapts as the generated distribution changes. This repeated training loop leaves inference as one evaluation of $\boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{c})$, without integrating the tracker field. Alg. 2 uses the same feedback structure with normalized attraction and, for $\lambda > 0$, generated-source repulsion. The displacement analogue illustrates this distributional correction mechanism; it is not the experimental update or a guarantee of exact pairwise transport.

2.5 Practical Deployment Guidance

Representation-space scattering. We use raw-pixel scattering for low-dimensional data and small images and scattering through frozen encoders for natural images. For a fixed representation map $\boldsymbol{f}$, Alg. 2 is applied to $\boldsymbol{f}(\mathbf{x}_\mathrm{p})$, $\boldsymbol{f}(\mathbf{x}_\mathrm{r})$, and $\boldsymbol{f}(\mathbf{x}_\mathrm{s})$; projectile gradients backpropagate through $\boldsymbol{f}$ into the generator, while source features remain detached. Pixel outputs enter $\boldsymbol{f}$ directly, whereas latent outputs are first decoded, so this projects the loss computation rather than changing the generator output domain. At $\lambda = 1$, the exact and denominator-smoothed objectives are proper only for the projected distributions; at $\lambda < 1$, the weighted functional is a projected surrogate. Neither case guarantees full image matching without sufficient representation informativeness.

Tracker parameterization. For ImageNet, each representation uses a four-layer conditional Transformer tracker with approximately 30M parameters and a vector-valued output head (App. B). This avoids the higher-order autograd of a scalar-potential tracker, while dropping exact conservativity.

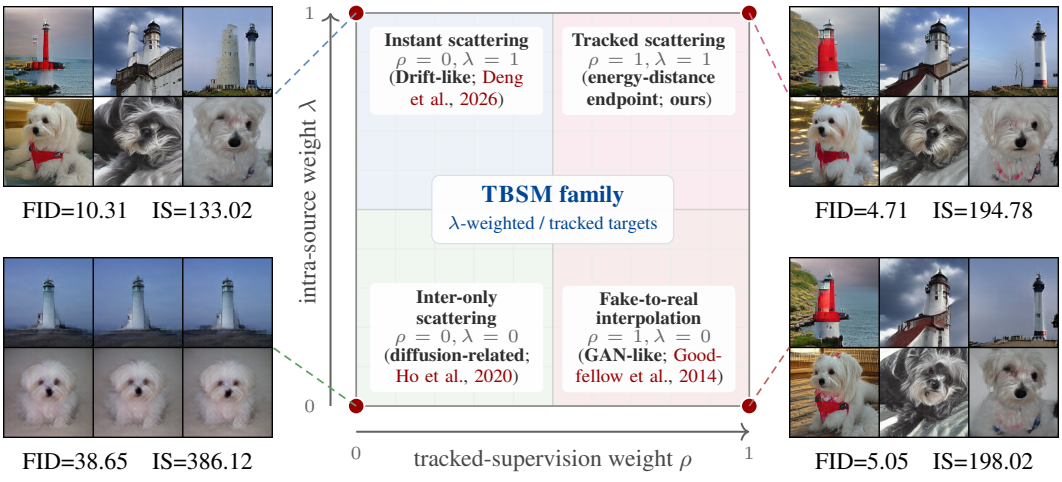


Figure 3: **One-step generated samples across the (ρ, λ) design map.** JiT-B, 5,000 generator updates, batch size 256; optimization settings are in Tab. 3.

The tracker and scattering feature extraction are used only during training; latent-output generators still require their decoder at inference.

Multiple representation fields. Each fixed representation map defines its own feature-space field, query, and regression loss; vectors from incompatible spaces are never added. At $\lambda = 1$, exact or smoothed, the expected current-parameter gradient of the summed instant or oracle-tracked losses equals that of the corresponding sum of proper projected discrepancies; finite-tracker errors combine only after Jacobian pullback. Zero summed discrepancy matches every projected marginal, not their joint or the full image law unless the representations are jointly measure-determining.

3 The Generative Design Map

Fig. 3 highlights two roles for the tracker. At $\lambda = 1$, tracking improves FID while its oracle preserves the energy-distance field, consistent with filtering source-sampling noise in an unchanged population direction (Cor. 1). This supports learning the correction field even when the instantaneous estimator is already unbiased.

More revealingly, the FID gap between $\lambda = 0$ and 1 is much smaller with tracking. Instantaneous attraction at $\lambda = 0$ favors contraction, whereas the oracle corridor field cancels at distributional agreement (Prop. 1). This suggests that distributional balance can enter supervision through explicit repulsion or through learning a corridor-conditioned fake-to-real field. The tracker query thus determines the expected correction as well as its noise. Because λ couples repulsion with query location, these corners do not isolate their contributions; diffusion- and GAN-like connections additionally require input or target changes (App. E).

4 Experiments

We evaluate one-step ImageNet generation and representation fields, with additional ImageNet studies (App. A) and Qwen-Image-20B benchmarks (App. C).

4.1 Experimental Setup

Datasets. Our primary quantitative evaluation uses ImageNet-1K (Deng et al., 2009) at 256×256 (ImageNet-256); App. A adds a 512×512 conversion study (ImageNet-512). MNIST, Fashion-MNIST, and CIFAR-10 are used only for qualitative observations in Fig. 1.

Table 1: **Representation-field ablation on ImageNet-256.** JiT-B, $(\rho, \lambda) = (0.5, 1)$, 5,000 generator updates, batch size 256. Runtime is per training step; optimizer settings are in Tab. 3.

Metric	ResNet-18	SigLIP2-B	MAE-B	ResNet-18 + SigLIP2-B	SigLIP2-B + MAE-B	ResNet-18 + MAE-B	ResNet-18 + MAE-B + SigLIP2-B
FID ↓	13.47	21.87	16.18	9.12	9.36	11.45	**8.29**
FDr^6 ↓	36.49	39.01	27.74	27.49	**21.80**	27.50	22.25
IS ↑	104.58	131.49	93.25	149.91	**161.36**	113.14	152.23
Avg. sec/step ↓	**0.1389**	0.1736	0.1583	0.1982	0.2184	0.1832	0.2411

Network architectures. On ImageNet, we evaluate JiT (Li & He, 2025), DiT (Peebles & Xie, 2023), and PixelDiT (Yu et al., 2025) backbones. JiT and PixelDiT operate directly in pixel space, while DiT generates latents that are decoded by a pretrained SD-VAE (Rombach et al., 2022). For the qualitative observations on MNIST, Fashion-MNIST, and CIFAR-10, we use compact U-Net generators. Tracker architectures and representation interfaces are detailed in App. B.

Implementation details. We use PyTorch (Paszke et al., 2019) and AdamW (Loshchilov & Hutter, 2017). Main ImageNet-256 runs use 100,000 updates for JiT-B (batch 256) and PixelDiT-XL (batch 128), and 50,000 for DiT-XL/2 (batch 256); Tab. 3 gives all training schedules. Except for the FD-loss refinement case, all ImageNet-1K runs initialize the generator from a pretrained multi-step diffusion or flow model. Checkpoint sources are listed in App. B. The instant baseline uses $\rho = 0, \lambda = 1$, with the unused tracker update omitted. All reported TBSM ImageNet metrics use 50,000 generated samples and training-set reference statistics. We report FID (Heusel et al., 2017), FDr^6 (Yang et al., 2026) (the mean generated-to-training FD normalized by validation-to-training FD over six feature spaces), and Inception Score (IS). NFE (number of function evaluations) counts the generator forward passes required to produce one sample. The corner comparison matches the generator-update budget, but tracked variants additionally optimize the tracker, so total training compute is not matched.

4.2 Representation-Field Ablation

Following FD-loss (Yang et al., 2026), we compare scattering vectors computed in single and multiple frozen feature spaces. The encoders include a supervised ResNet-18 (He et al., 2016), SigLIP2-B (Tschannen et al., 2025), and MAE-B (He et al., 2022). For each encoder, we EMA-whiten and channelwise ℓ_2-normalize its final-layer map, then use a separate tracker and equally weighted loss (App. B). Tab. 1 is the only ImageNet-1K experiment that varies this set; every other ImageNet-1K experiment sums all three losses.

Tab. 1 suggests that an encoder's value depends on the other fields used alongside it. SigLIP2-B has the weakest FID when used alone, yet both pairs containing it outperform every single encoder on all three quality metrics. This is consistent with different representations exposing complementary distributional discrepancies to the generator, although each additional encoder also introduces tracker capacity and computation. Adding ResNet-18 to SigLIP2-B + MAE-B further improves FID but worsens FDr^6 and IS. We use all three fields for the best FID in this ablation; the two-encoder combination offers lower training cost and stronger scores on the other two metrics.

4.3 Benchmark Results

Tab. 2 shows competitive one-step generation with both pixel and latent outputs, from the smaller JiT-B to the larger PixelDiT and DiT backbones. Since scattering is computed in image-feature spaces in each case, the same supervision interface accommodates these different generator parameterizations. FD-loss (Yang et al., 2026) retains the strongest table scores.

At fixed $\rho = 0.9$, lowering λ from 1 to 0.9 raises IS on both pixel backbones but worsens FDr^6. FID improves on JiT-B and is unchanged to reported precision on PixelDiT-XL. Thus the improvement under FID and IS does not extend to closer agreement across the feature spaces used by FDr^6. This resembles the metric tradeoffs associated with guidance (Ho & Salimans, 2022), although λ acts through source weighting and corridor queries.

Table 2: **ImageNet benchmark comparison.** Space denotes the generator output domain, and Params counts inference-time generator parameters. TBSM rows show their (ρ, λ) settings; † denotes longer training.

Method	Space	Backbone	NFE	Params	FID ↓	FDr^6 ↓	IS ↑
Reference statistics							
Validation set images	pixel	-	-	-	1.68	1.00	232.2
Iterative diffusion and flow baselines							
ADM (Dhariwal & Nichol, 2021)	pixel	U-Net cascade	250+250	608M	3.94	–	215.8
VDM++ (Kingma & Gao, 2023)	pixel	U-ViT-L/2	512×2	2B	2.12	–	267.7
DiT (Peebles & Xie, 2023)	latent	DiT-XL/2	250×2	675M	2.27	–	278.2
JiT (Li & He, 2025)	pixel	JiT-H/16	200	953M	1.97	7.66	296.0
PixelDiT (Yu et al., 2025)	pixel	PixelDiT-XL/16	100×2	797M	1.61	–	292.7
Autoregressive visual generators							
VAR (Tian et al., 2024)	discrete	VAR-d30	10×2	2B	1.97	6.70	304.6
BAR (Yu et al., 2026)	discrete	BAR-L	2048	1.1B	1.01	3.57	281.9
MAR (Li et al., 2024)	latent	MAR-H	51,200	942M	1.56	5.61	299.5
GAN-based one-step generators							
BigGAN (Brock et al., 2018)	pixel	BigGAN-deep	1	–	6.95	–	171.4
StyleGAN (Sauer et al., 2022)	pixel	StyleGAN-XL	1	–	2.30	–	265.1
One-step and few-step diffusion acceleration							
iCT (Song & Dhariwal, 2023)	latent	DiT-XL/2	1	675M	34.24	–	–
Shortcut (Frans et al., 2024)	latent	DiT-XL/2	1	676M	10.60	–	–
MeanFlow (Geng et al., 2025)	latent	DiT-XL/2	1	676M	3.43	–	–
pMF (Lu et al., 2026)	pixel	pMF-H/16	1	935M	2.29	6.87	267.2
Direct distribution-dynamics and representation-distance methods							
Drift (Deng et al., 2026)	pixel	DiT-L/16	1	465M	1.43	10.51	305.8
FD-loss (Yang et al., 2026)	pixel	JiT-H/16	1	953M	0.75	2.65	313.0
TBSM variants							
TBSM (ρ=1.0, λ=1.0)	pixel	JiT-B/16	1	131M	3.09	10.73	208.6
TBSM (ρ=0.9, λ=1.0)	pixel	JiT-B/16	1	131M	3.35	11.60	205.3
TBSM (ρ=0.9, λ=0.9)	pixel	JiT-B/16	1	131M	2.92	13.21	228.7
TBSM† (ρ=0.9, λ=0.9)	pixel	JiT-B/16	1	131M	2.69	13.21	247.4
TBSM (ρ=0.9, λ=1.0)	pixel	PixelDiT-XL/16	1	797M	2.23	7.15	210.0
TBSM (ρ=0.9, λ=0.9)	pixel	PixelDiT-XL/16	1	797M	2.23	7.59	230.9
TBSM (ρ=0.9, λ=0.9)	latent	DiT-XL/2	1	675M	1.63	6.76	243.0

The ImageNet-512 study illustrates why the pretrained initialization still needs to be adapted for one-step sampling. The original DiT-XL/4 produces poor samples in a single pass, whereas TBSM training makes the same architecture a successful one-step generator (App. A).

Text-to-image transfer. The Qwen-Image-20B study applies the same learning loop to individual text–image pairs, where each loss requires only one real image for its caption. After 1,000 updates at $(\rho, \lambda) = (1, 0)$, the pretrained model produces one-step samples at 1024×1024 resolution (Fig. 1). It achieves GenEval 0.88, DPG-Bench 86.34, and WISE 0.56 (App. C), providing quantitative evidence of transfer; conditional diversity remains unassessed.

5 Discussion and Conclusion

We presented TBSM, which learns a fake-to-real correction field through an online tracker and transfers its predictions into a one-step generator through frozen-target regression. Our analysis explains when tracking filters persistent interaction noise and establishes identifiability of the smoothed full-corridor field.

The tracker can accumulate information across training events, but every generator update also changes the field it must learn. At $\lambda = 1$, increasing ρ reduces source-sampling noise while placing more weight on tracker accuracy (Cor. 1), making adaptation to the current generator part of the estimation problem. Further limitations are discussed in App. G.

AI USE STATEMENT

Generative AI tools assisted with manuscript editing, related-work literature review, selected theoretical derivations and numerical checks. The authors manually reviewed the AI-assisted content and take responsibility for all claims, proofs, code, and results in this work.

REFERENCES

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser, 2 edition, 2008.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Marc G. Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The Cramér distance as a solution to biased Wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.

Siwan Boufadène and François-Xavier Vialard. On the global convergence of Wasserstein gradient flow of the Coulomb discrepancy. *SIAM Journal on Mathematical Analysis*, 57(4):4556–4587, 2025.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Paul Caucheteux, Clément Bonet, and Anna Korba. A unifying view of variational generative wasserstein flows. *arXiv preprint arXiv:2605.31369*, 2026.

Zhenglin Cheng, Peng Sun, Jianguo Li, and Tao Lin. TwinFlow: Realizing one-step generation on large models with self-adversarial flows. *arXiv preprint arXiv:2512.05150*, 2025.

Lénaïc Chizat, Maria Colombo, Roberto Colombo, and Xavier Fernández-Real. Quantitative convergence of Wasserstein gradient flows of kernel mean discrepancies. *arXiv preprint arXiv:2603.01977*, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Richard Duong, Viktor Stein, Robert Beinert, Johannes Hertrich, and Gabriele Steidl. Wasserstein gradient flows of MMD functionals with distance kernel and cauchy problems on quantile functions. *ESAIM: Control, Optimisation and Calculus of Variations*, 32:10, 2026. doi: 10.1051/cocv/2025097.

Maria Esteban-Casadevall, Jorge Carrasco-Pollo, Max Welling, Jan-Willem van de Meent, Erik J. Bekkers, and Floor Eijkelboom. Kernel-gradient drifting models. *arXiv preprint arXiv:2605.10727*, 2026.

Lan Feng, Wuyang Li, Eloi Zablocki, Matthieu Cord, and Alexandre Alahi. Representation distribution matching for one-step visual generation. *arXiv preprint arXiv:2607.02375*, 2026.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. *arXiv preprint arXiv:2410.12557*, 2024.

Leonard T Franz, Sebastian Hoffmann, Tim Weiland, Bernhard Schölkopf, and Georg Martius. Drifting fields are not conservative. *arXiv preprint arXiv:2604.06333*, 2026.

Alexandre Galashov, Valentin de Bortoli, and Arthur Gretton. Deep MMD gradient flow without adversarial training. In *International Conference on Learning Representations*, 2025.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. *arXiv preprint arXiv:2310.11513*, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Paul Hagemann, Johannes Hertrich, Fabian Altekrüger, Robert Beinert, Jannis Chemseddine, and Gabriele Steidl. Posterior sampling based on gradient flows of the MMD with negative distance kernel. In *International Conference on Learning Representations*, 2024.

Jiaqi Han, Puheng Li, Qiushan Guo, Renyuan Xu, Stefano Ermon, and Emmanuel J. Candès. One-step generative modeling via wasserstein gradient flows. *arXiv preprint arXiv:2605.11755*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Ping He, Om Khangaonkar, Hamed Pirsiavash, Yikun Bai, and Soheil Kolouri. Sinkhorn-drifting generative models. *arXiv preprint arXiv:2603.12366*, 2026.

Johannes Hertrich, Christian Wald, Fabian Altekrüger, and Paul Hagemann. Generative sliced MMD flows with Riesz kernels. In *International Conference on Learning Representations*, 2024.

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. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10124–10134, 2023.

Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Advances in Neural Information Processing Systems*, 2023.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024.

Yujia Li, Kevin Swersky, and Richard Zemel. Generative moment matching networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1718–1727, 2015.

Shanchuan Lin, Ceyuan Yang, Zhijie Lin, Hao Chen, and Haoqi Fan. Adversarial flow models. *arXiv preprint arXiv:2511.22475*, 2025.

Shanchuan Lin, Ceyuan Yang, Zhijie Lin, Hao Chen, and Haoqi Fan. Continuous adversarial flow models. *arXiv preprint arXiv:2604.11521*, 2026.

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

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. *arXiv preprint arXiv:2410.11081*, 2024.

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows. *arXiv preprint arXiv:2601.22158*, 2026.

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*, pp. 23–40. Springer, 2024.

Zehong Ma, Ruihan Xu, and Shiliang Zhang. PixelGen: Improving pixel diffusion with perceptual supervision. *arXiv preprint arXiv:2602.02493*, 2026.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Fanqing Meng, Kunpeng Ning, Bin Zhu, and Li Yuan. WISE: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Nicolaj Rux, Michael Quellmalz, and Gabriele Steidl. Smoothed distance kernels for MMDs and applications in Wasserstein gradient flows. *Advances in Computational Mathematics*, 52:24, 2026.

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

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

René L. Schilling, Renming Song, and Zoran Vondraček. *Bernstein Functions: Theory and Applications*. De Gruyter, 2 edition, 2012. doi: 10.1515/9783110269338.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013.

Daniil Shlenskii, Nikita Gushchin, Lev Novitskiy, Dmitry V. Dylov, and Alexander Korotin. Midpoint generative models. *arXiv preprint arXiv:2605.29920*, 2026.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. *arXiv preprint arXiv:2310.14189*, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

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

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Peng Sun, Yi Jiang, and Tao Lin. Unified continuous generative models. *arXiv preprint arXiv:2505.07447*, 2025.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

Chenyang Tian, Bharath K. Sriperumbudur, Arthur Gretton, and Zonghao Chen. Sobolev regularized MMD gradient flow. *arXiv preprint arXiv:2605.11884*, 2026.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Thomas Unterthiner, Bernhard Nessler, Calvin Seward, Günter Klambauer, Martin Heusel, Hubert Ramsauer, and Sepp Hochreiter. Coulomb GANs: Provably optimal Nash equilibria via potential fields. In *International Conference on Learning Representations*, 2018.

Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1747–1756, 2016.

Yakun Wang, Leyang Wang, Song Liu, and Taiji Suzuki. Zero-flow encoders. *arXiv preprint arXiv:2602.00797*, 2026.

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, pp. 8406–8441, 2023.

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

Jiawei Yang, Zhengyang Geng, Xuan Ju, Yonglong Tian, and Yue Wang. Representation Fréchet loss for visual generation. *arXiv preprint arXiv:2604.28190*, 2026.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *arXiv preprint arXiv:2405.14867*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024b.

Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

Qihang Yu, Qihao Liu, Ju He, Xinyang Zhang, Yang Liu, Liang-Chieh Chen, and Xi Chen. Autoregressive image generation with masked bit modeling. *arXiv preprint arXiv:2602.09024*, 2026.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. PixelDiT: Pixel diffusion transformers for image generation. *arXiv preprint arXiv:2511.20645*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023.

Chuyang Zhao, Yifei Song, Hongfa Wang, Jianlong Yuan, Yuan Zhang, Siming Fu, Zhineng Chen, Huilin Deng, Haoyang Huang, and Nan Duan. Perceptual flow matching for few-step generative modeling. *arXiv preprint arXiv:2607.03524*, 2026.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 62307–62331, 2024.

CONTENTS

1 Introduction **2**

2 Methodology **3**
2.1 Distributional Objective and Stochastic Optimization . . . 3
2.2 Three-Body Scattering Vector Field . . . 4
2.3 Scattering Target Regression . . . 4
2.4 Tracked Scattering Targets . . . 5
2.5 Practical Deployment Guidance . . . 6

3 The Generative Design Map **7**

4 Experiments **7**
4.1 Experimental Setup . . . 7
4.2 Representation-Field Ablation . . . 8
4.3 Benchmark Results . . . 8

5 Discussion and Conclusion **9**

A Additional ImageNet Results **16**

B Implementation Details **17**

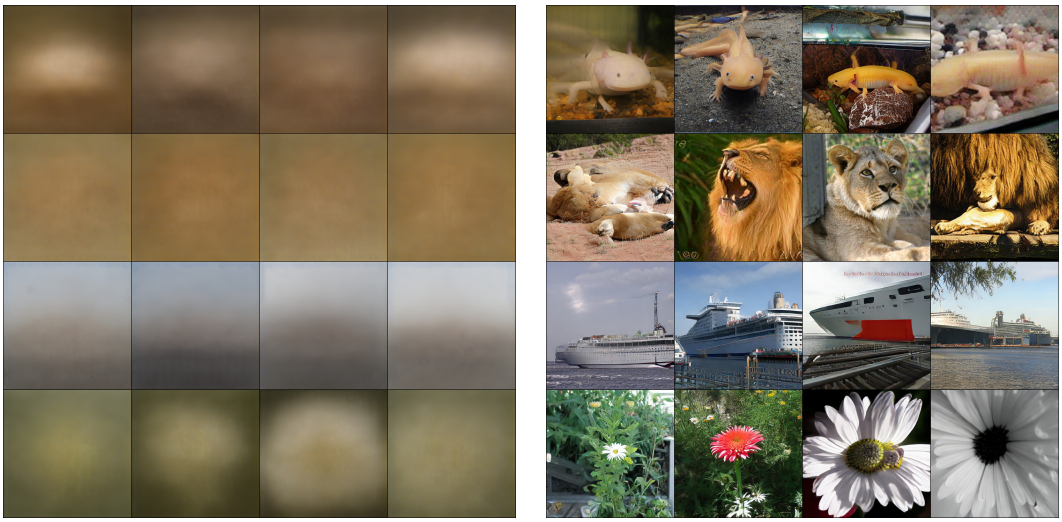
C Text-to-Image Generation **19**

D Related Work and Positioning **20**
D.1 Adversarial One-Step Generators . . . 20
D.2 Diffusion, Score, and Flow-Based Models . . . 20
D.3 Autoregressive Visual Generation . . . 20
D.4 One-Step and Few-Step Diffusion Acceleration . . . 20
D.5 Generative Wasserstein Flows and Particle Dynamics . . . 21

E Design-Map Interpretations **22**
E.1 Displacement Analogue of the GAN-Like State . . . 22
E.2 Reference Configurations and Matched Generator Updates . . . 23
E.3 Connections to Related Distribution-Matching Objectives . . . 23

F Derivation and Theoretical Foundations **24**
F.1 Exact Scattering Field and Frozen-Target Equivalence . . . 24
F.2 Why Tracking Helps: Noise, Projection, and Identifiability . . . 25
F.2.1 Nonvanishing interactions at distributional agreement . . . 25
F.2.2 Conditional-expectation projection and finite tracker error . . . 26
F.2.3 Identifiability of the full-corridor tracker field . . . 27
F.2.4 Local signal-to-noise in a Gaussian location model . . . 28
F.3 Population Dissipation and Idealized Optimization Guarantees . . . 29
F.3.1 Exact-flow dissipation . . . 29
F.3.2 Tracked population flow . . . 30
F.3.3 Finite-generator reference recursion . . . 31
F.4 Source-Law Necessity and Variance Optimality . . . 32
F.5 Implementation and Interpretive Derivations . . . 33
F.5.1 Denominator smoothing as a proper objective . . . 33
F.5.2 Near-endpoint surrogate control . . . 33
F.5.3 Line-integral interpretation of the displacement tracker . . . 35

G Limitations **35**



(a) FID = 398.20 FDr^6 = 291.50 IS = 1.17. (b) FID = 1.92 FDr^6 = 5.37 IS = 252.1.

Figure 4: **Converting a multi-step ImageNet-512 model into a one-step generator.** Single-pass samples from a pretrained multi-step DiT-XL/4 before (left) and after TBSM training with $\rho = \lambda = 0.9$ (right). Both grids are random and uncurated.



(a) FID = 0.99 FDr^6 = 5.64 IS = 328.0. (b) FID = 1.84 FDr^6 = 8.34 IS = 247.3.

Figure 5: **FD-loss-trained JiT-B checkpoint refinement on ImageNet-256.** Random, uncurated single-pass samples before (left) and after continued TBSM training with $\rho = \lambda = 0.9$ (right).

A Additional ImageNet Results

Tab. 3 gives optimizer, batch-size, and training-step settings for both studies.

Multi-step-to-one-step conversion at 512×512. The left panel of Fig. 4 evaluates a pretrained multi-step DiT-XL/4 at NFE $= 1$, outside its intended sampler; after TBSM training, the same architecture operates as an NFE $= 1$ ImageNet-512 generator. For the first half of this high-resolution experiment, we use an empirical paired noisy-input curriculum with Alg. 2: we sample $t \in \{0.5, 1.0\}$ with equal probability, form $\mathbf{x}_t = t\mathbf{z} + (1-t)\mathbf{x}_\mathrm{r}$, set $\mathbf{x}_\mathrm{p} = \boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \mathbf{c})$, and additionally condition the tracker on t. Throughout, the generated source $\mathbf{x}_\mathrm{s} = \mathrm{sg}[\boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}_\mathrm{s}, 1, \mathbf{c})]$ uses an independent pure-noise draw $\mathbf{z}_\mathrm{s}$ and does not depend on the sampled t. When $t = 0.5$, however, $\mathbf{x}_\mathrm{p}$ is constructed from $\mathbf{x}_\mathrm{r}$ and therefore does not satisfy real-source–projectile independence; moreover, $\mathbf{x}_\mathrm{p}$ and $\mathbf{x}_\mathrm{s}$ are not

Table 3: **ImageNet training configurations: shared optimizer settings and per-experiment batch sizes and steps.**

Configuration	Shared setting
Generator optimizer	AdamW
Generator learning rate	1.0×10^{-5}
Generator betas	$(0.9, 0.999)$
Tracker optimizer	AdamW
Tracker learning rate	1.0×10^{-3}
Tracker betas	$(0.9, 0.999)$
Generator EMA decay	0.999
Training precision	BF16

Configuration	JiT-B (Tab. 2)	PixelDiT-XL (Tab. 2)	DiT-XL/2 (Tab. 2)	DiT-XL/4 (Fig. 4)	JiT-B (Tab. 1; Fig. 3, Fig. 5)
Global batch size	256	128	256	256	256
Training steps	100,000 (200,000 for †)	100,000	50,000	25,000	5,000

sampled iid from the same current generated law, as assumed by the exact field and convergence analysis. We use this curriculum solely as an empirical stabilization heuristic; none of our theoretical claims relies on it. For the second half, we fix $t = 1.0$, restoring the pure-noise generator input used at one-step deployment. Using the same optimization settings, PixelDiT-XL reaches FID $= 3.84$, IS $= 236.44$, and FDr^6 $= 8.17$ on ImageNet-512.

FD-loss checkpoint case study. In this separate JiT-B case, the post-training grid displays fewer block-like and grid-like artifacts although all three aggregate metrics worsen (Fig. 5; Yang et al., 2026). This illustrates a possible metric–artifact mismatch in the displayed samples, not a general limitation of FD-loss.

B Implementation Details

Each training step performs one generator update and, when $\rho > 0$, one tracker update on the same triplet minibatch. The representation-field ablation in Tab. 1 fixes $(\rho, \lambda) = (0.5, 1)$ for all encoder combinations.

Numerical stabilization. All reported experiments compute each bearing as $(\mathbf{y}-\mathbf{x})/(\|\mathbf{y}-\mathbf{x}\|_2+\varepsilon)$ with fixed, untuned $\varepsilon = 10^{-6}$ to prevent division by zero. Sec. F.5.1 characterizes the induced smoothed objective and its relation to the exact bearing field.

Exact endpoint and deployed surrogates. At $\lambda = 1$, exact and smoothed source means are the negative Wasserstein-gradient fields of the proper objectives $\mathcal{F}$ and $\mathcal{F}_\varepsilon$ (App. F.5.1). Tracker identities apply to both; tracked convergence additionally requires the corresponding chain rule, slope, and relative-error conditions. For $\lambda < 1$, Prop. 7 gives objective consistency as $\lambda \uparrow 1$, while Prop. 8 combines reweighting, corridor, and tracker errors; under a linear endpoint slope, these errors yield an explicit convergence neighborhood.

Tracker architecture. The tracker is a lightweight conditional Transformer with approximately 30M parameters, comprising four layers with hidden dimension 736, eight attention heads, and an MLP expansion ratio of 4. Input features are linearly embedded and augmented with two-dimensional sinusoidal positional encodings and a joint class–timestep conditioning embedding, then processed by four self-attention–MLP blocks with RMSNorm, residual connections, and spectral normalization. Reported image experiments use a linear vector-valued head that projects the resulting tokens back to the input feature dimension. The conservative-field alternative instead predicts a scalar potential per token; these potentials are summed, and differentiating the resulting scalar with respect to the input yields the vector field.

Representation interfaces. For latent-output models, each generated latent is decoded by the corresponding frozen decoder before the resulting image is passed through the frozen representation encoders used to compute the scattering loss; ImageNet DiT uses the SD-VAE. On the projectile



Figure 6: **One-step text-to-image generation.** NFE $= 1$ samples generated at 1024×1024 resolution by a TBSM-fine-tuned Qwen-Image-20B model.

branch, gradients propagate through both decoder and encoders into the generator, while source features remain detached as in Alg. 2. Thus, "latent" and "pixel" in Tab. 2 denote the generator output domain; scattering is computed in frozen image-representation spaces in both cases. For every representation model, we use only its final-layer feature map; unlike Drifting Models (Deng et al., 2026), which combines features from multiple stages, scales, and locations, we do not aggregate intermediate layers. For a feature tensor $\boldsymbol{F} \in \mathbb{R}^{B\times C\times H\times W}$, each minibatch supplies channel-wise means and standard deviations over (B, H, W) in $\mathbb{R}^{1\times C\times 1\times 1}$; we maintain detached exponential moving averages with decay 0.999. Each sample is whitened with the pre-update running statistics and then ℓ_2-normalized along C at each spatial location. For scattering, we then identify each normalized $C \times H \times W$ sample with its vectorization in $\mathbb{R}^{CHW}$, so every bearing denominator uses the global Euclidean norm (equivalently, the tensor Frobenius norm) across all channels and spatial locations, not a positionwise norm. Only after the losses are formed are the running statistics updated with the current minibatch. Thus the preprocessing map is fixed conditional on the history before each loss event, so the current-step estimator and tracker identities remain valid conditionally on that history. The theory below treats the vectors entering the scattering objective as the primitive sample space and does not cover the slow time variation induced by these EMA updates. Each representation model has its own tracker; all representation branches use the same feature processing and tracker architecture, and their losses are summed only after being formed in their respective feature spaces.

Generator initialization. Except for the FD-loss checkpoint case in Fig. 5, all ImageNet-1K TBSM runs initialize the generator from an architecture- and resolution-matched multi-step checkpoint trained with a diffusion or flow-matching objective. For JiT and PixelDiT, we use the authors' public checkpoints for their best-reported multi-step configurations (Li & He, 2025; Yu et al., 2025), available from the official project repositories. For DiT, we instead initialize DiT-XL/2 and DiT-XL/4 from the multi-step checkpoints released by UCGM (Sun et al., 2025). The initialized network is then optimized with the relevant setting of Alg. 2. Unlike consistency distillation (Song et al., 2023), which couples predictions across noise levels, and DMD (Yin et al., 2024b), which queries

diffusion and generated-distribution scores, this warm start supplies only initial parameters: no consistency target, teacher prediction, score, or noise-time target enters the TBSM update. Thus, the reported one-step models use TBSM supervision and a single-pass sampler; "one-step" does not imply random initialization. For the primary runs, this practical warm start improves training efficiency while leaving the TBSM objective independent of the initialization source. Tab. 3 reports the shared optimizer settings and per-experiment batch sizes and step counts for this TBSM training.

C TEXT-TO-IMAGE GENERATION

Initialization and supervision. We initialize the generator directly from the original pretrained multi-step Qwen-Image-20B (Wu et al., 2025) checkpoint, without an intermediate distillation stage or teacher-assisted conversion. Thereafter, all generator and tracker updates use only Alg. 2. Relative to teacher-based distillation and adversarial post-training pipelines (Yin et al., 2024b; Lin et al., 2026), the supervision interface is simpler in this respect: it requires no path-model teacher query, score estimate, denoising-trajectory target, or adversarial critic.

Model and representation interfaces. The samples in Fig. 1 and Fig. 6 are generated at NFE $= 1$ without classifier-free guidance. Relative to the ImageNet implementation, each tracker introduces cross-attention to the text-conditioning sequence from the Qwen2.5-VL (Bai et al., 2025) text encoder in Qwen-Image-20B. We form separate scattering losses from the final-layer features of three frozen representation models: the same Qwen2.5-VL encoder, DINOv3 (Siméoni et al., 2025), and SigLIP (Zhai et al., 2023). Each representation has its own tracker and follows the whitening, normalization, and loss-aggregation procedure in Sec. B.

Training configuration. Our training-data setup follows the Qwen-Image-20B experiments in Table 3 of TwinFlow (Cheng et al., 2025). We set $\rho = 1$ and $\lambda = 0$ and train for 1,000 generator updates with batch size 16. The generator uses learning rate 5×10^{-6} and AdamW betas $(0.9, 0.99)$; each tracker uses learning rate 1×10^{-5} and betas $(0.9, 0.999)$. Training likewise uses no classifier-free guidance. Tab. 4 reports our NFE $= 1$ scores on GenEval (Ghosh et al., 2023), DPG-Bench (Hu et al., 2024), and WISE (Niu et al., 2025) alongside baseline results.

Table 4: **Text-to-image generation with Qwen-Image-20B.** Higher scores are better on all three benchmarks.

Method	NFE↓	GenEval↑	DPG-Bench↑	WISE↑
Qwen-Image (Wu et al., 2025)	50×2	0.87	88.32	0.62
VSD (Wang et al., 2023)	1	0.67	84.44	0.22
	2	0.73	86.16	0.34
DMD (Yin et al., 2024b)	1	0.81	84.31	0.47
	2	0.80	84.08	0.46
SiD (Zhou et al., 2024)	1	0.77	87.05	0.42
	2	0.78	86.94	0.41
sCM (JVP-free) (Lu & Song, 2024)	4	0.62	85.37	0.44
	8	0.60	85.54	0.45
MeanFlow (JVP-free) (Geng et al., 2025)	4	0.44	83.28	0.34
	8	0.49	83.81	0.37
TwinFlow (Cheng et al., 2025)	1	0.85	85.44	0.51
	2	0.86	86.35	0.55
TBSM (ours)	1	0.88	86.34	0.56

VSD, DMD, and SiD use LoRA fake-score estimators with rank 64. For sCM and MeanFlow, JVPs are approximated by finite differences. Qwen-Image uses classifier-free guidance (50×2 NFE).

Table 5: **Positioning of major generative modeling paradigms.** We compare direct single-pass sampling, the breadth of empirical scaling evidence, and the granularity of generator supervision.

Paradigm / representative work	Direct single-pass sampling?	Scaling evidence?	Generator-supervision granularity
GANs (Goodfellow et al., 2014; Brock et al., 2018; Sauer et al., 2022; Kang et al., 2023)	**Yes** direct single-pass generation	**Mixed** several large-scale successes	**Per-sample** signal from a learned critic
Diffusion, score, and flow models (Ho et al., 2020; Song et al., 2020b; Lipman et al., 2022; Ma et al., 2024)	**No** iterative path-based sampling	**Strong** extensive evidence at large scale	**Per-sample** denoising, score, or velocity target
Autoregressive models (van den Oord et al., 2016; Tian et al., 2024; Sun et al., 2024)	**No** iterative structured decoding	**Strong** extensive evidence at large scale	**Per-token** predictive likelihood target
One-step or few-step diffusion acceleration (Song et al., 2023; Frans et al., 2024; Geng et al., 2025; Yin et al., 2024b)	**Method-dependent** one or a few generator evaluations	**Growing** demonstrated on large image models	**Method-dependent** teacher, consistency, or auxiliary target
Direct distribution dynamics and distributional losses (Deng et al., 2026; Yang et al., 2026; Caucheteux et al., 2026)	**Yes** direct single-pass generation	**Growing** recent ImageNet-scale results	**Batch-estimated** fields, statistics, or objectives

D Related Work and Positioning

We position TBSM from its exact energy-distance endpoint: an explicit distributional energy whose population field is estimated with constant-size stochastic interactions, without adversarial, path-indexed, autoregressive, or distillation-based supervision.

D.1 Adversarial One-Step Generators

GANs (Goodfellow et al., 2014; Brock et al., 2018; Sauer et al., 2022; Kang et al., 2023) provide direct single-pass sampling but train the generator through a jointly learned discriminator or critic. Adversarial Flow Models use adversarial training for native one- or multi-step flow-style generation, while Continuous Adversarial Flow Models adversarially post-train continuous flow models (Lin et al., 2025; 2026). TBSM instead uses non-adversarial frozen-target regression; at its exact endpoint, the sampled scattering vector estimates the 2-Wasserstein negative-gradient field of the energy-distance functional. A GAN-like objective arises only for the conservative displacement analogue in Sec. E.1, where (13) gives a quadratically path-regularized potential objective.

D.2 Diffusion, Score, and Flow-Based Models

Standard diffusion, score-based, and flow-matching models (Ho et al., 2020; Song et al., 2020b;a; Lipman et al., 2022; Ma et al., 2024) use path-indexed supervision: a network predicts a denoising target, score, or velocity along a noise-to-data trajectory and is typically evaluated repeatedly during sampling. The exact energy-distance formulation requires no noise schedule or transport path; it updates the generator distribution directly through real-generated-generated particle interactions.

D.3 Autoregressive Visual Generation

Autoregressive visual generators (van den Oord et al., 2016; Yu et al., 2023; Li et al., 2024; Tian et al., 2024; Sun et al., 2024) formulate image generation as ordered conditional prediction over discrete or continuous token, patch, or scale representations. They exploit sequence-modeling infrastructure, while sampling follows a chosen order or scale schedule; parallel and masked variants can reduce the number of decoding rounds. TBSM is non-autoregressive: it emits a full continuous sample in one generator evaluation and trains it through the population scattering vector field.

D.4 One-Step and Few-Step Diffusion Acceleration

The sampling cost of diffusion and flow models has motivated a large body of one-step and few-step acceleration methods, including progressive distillation (Salimans & Ho, 2022), consistency

models (Song et al., 2023; Lu & Song, 2024; Song & Dhariwal, 2023), shortcut models (Frans et al., 2024), mean flows (Geng et al., 2025), and distribution matching distillation (Yin et al., 2024b;a). They obtain one- or few-step sampling through method-specific combinations of path- or noise-time supervision, consistency constraints, teachers, or auxiliary distribution estimators. At its energy-distance endpoint, TBSM instead trains the one-step generator with a population distribution-matching objective that does not require teacher outputs, trajectory targets, or score estimates. Our primary ImageNet experiments use a multi-step checkpoint only for parameter initialization, as detailed in Sec. B; this practical warm start is distinct from making a path model part of the TBSM supervision.

D.5 Generative Wasserstein Flows and Particle Dynamics

The closest line directly optimizes distributional discrepancies or their induced dynamics. Energy distance is equivalent to MMD with a distance-induced kernel (Sejdinovic et al., 2013); both yield proper population objectives under their standard assumptions (Székely & Rizzo, 2013; Gretton et al., 2012). Generative Moment Matching Networks directly backpropagate an MMD objective (Li et al., 2015), whereas MMD-GAN and Cramér GAN learn the comparison kernel or feature map adversarially (Li et al., 2017; Bellemare et al., 2017). Coulomb GAN uses a learned potential field with data attraction and generated-particle repulsion (Unterthiner et al., 2018). Under the normalization below, the $\lambda = 1$ population field in (3) is exactly the velocity field of the 2-Wasserstein gradient flow of distance-kernel MMD^2; alternative constant normalizations only rescale time. MMD Wasserstein flows have been developed for interacting-particle transport, one-dimensional negative-distance dynamics, sliced Riesz kernels, posterior sampling, and deep generative modeling (Arbel et al., 2019; Duong et al., 2026; Hertrich et al., 2024; Hagemann et al., 2024; Galashov et al., 2025). Smoothed distance kernels have also been constructed to restore differentiability and flow well-posedness near particle collisions (Rux et al., 2026). Recent work strengthens this theory in complementary directions: Sobolev-regularized MMD flow penalizes witness gradients and proves global MMD convergence under embedding-regularity assumptions, while kernel-gradient drifting expresses general-kernel drifts as score differences between kernel-smoothed laws, yielding a smoothed-KL descent interpretation (Tian et al., 2026; Esteban-Casadevall et al., 2026). The population energy flow is therefore not claimed as new; our theory concerns its constant-size source-detached estimator, online conditional-expectation tracking, and finite-error transfer to generator updates. More broadly, Wasserstein gradient-flow theory and Stein variational methods provide related distributional and interacting-particle views (Jordan et al., 1998; Ambrosio et al., 2008; Liu & Wang, 2016). MMD admits both full pairwise empirical estimates and linear-time random-pair approximations (Gretton et al., 2012).

Relation to linear-time MMD estimation. For distributions with finite first moments and any anchor $\mathbf{a}$, the distance-induced kernel $k_{\mathbf{a}}(\mathbf{x}, \mathbf{y}) = \frac{1}{2}(\|\mathbf{x} - \mathbf{a}\| + \|\mathbf{y} - \mathbf{a}\| - \|\mathbf{x} - \mathbf{y}\|)$ makes $\frac{1}{2} D_E^2(P_{\boldsymbol{\theta}}, Q)$ equal to the squared MMD. For independent $\mathbf{x}_1, \mathbf{x}_2 \sim P_{\boldsymbol{\theta}}$ and $\mathbf{x}_{\mathrm{r},1}, \mathbf{x}_{\mathrm{r},2} \sim Q$, one block of its classical linear-time estimator and its generator gradient can be written as

$$\begin{aligned} \widehat{\mathcal{F}}_{\mathrm{lin}} &= \frac{1}{2}\left(\|\mathbf{x}_1 - \mathbf{x}_{\mathrm{r},2}\| + \|\mathbf{x}_2 - \mathbf{x}_{\mathrm{r},1}\| - \|\mathbf{x}_1 - \mathbf{x}_2\| - \|\mathbf{x}_{\mathrm{r},1} - \mathbf{x}_{\mathrm{r},2}\|\right), \\ \nabla_{\boldsymbol{\theta}} \widehat{\mathcal{F}}_{\mathrm{lin}} &= -\frac{1}{2}\left(J_1^\top \widehat{\mathbf{v}}_{\mathrm{scat}}(\mathbf{x}_1; \mathbf{x}_{\mathrm{r},2}, \mathbf{x}_2) + J_2^\top \widehat{\mathbf{v}}_{\mathrm{scat}}(\mathbf{x}_2; \mathbf{x}_{\mathrm{r},1}, \mathbf{x}_1)\right), \end{aligned} \tag{12}$$

where $J_i = \nabla_{\boldsymbol{\theta}} \mathbf{x}_i$ and the identity holds away from coincident samples. Thus a symmetrized linear-time distance-kernel MMD gradient decomposes into two three-body projectile updates. By exchangeability, sampling either projectile term uniformly and omitting its factor $1/2$ has the same expectation as the displayed average. Alg. 1 realizes this update through regression with the generated source detached; up to conventional loss scaling, Thm. 1 proves equality with the energy-distance gradient at the current parameters. Thus Alg. 1 is a source-detached projectile-level realization of a classical stochastic gradient, not a new linear-time MMD estimator.

Recent direct distribution-training methods differ mainly in how they estimate distribution-level supervision. Drifting Models construct minibatch kernel attraction–repulsion fields, with subsequent work analyzing when such fields are conservative (Deng et al., 2026; Franz et al., 2026); FD-loss instead optimizes representation-space feature means and covariances using population statistics and batch-level gradients (Yang et al., 2026). Representation Distribution Matching uses a frozen Nyström

Algorithm 3 GAN-like Displacement Analogue of TBSM at $\rho = 1, \lambda = 0$

1: Initialize generator parameters $\boldsymbol{\theta}$ and tracker parameters $\boldsymbol{\phi}$.
2: **repeat**
3: Sample real source-condition pairs $(\mathbf{x}_\mathrm{r}, \mathbf{c})$ from dataset.
4: Sample noise $\mathbf{z} \sim p(\mathbf{z})$ and generate projectile $\mathbf{x}_\mathrm{p} = \boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{c})$.
5: Sample $\alpha \sim \mathcal{U}(0, 1)$ and set $\tilde{\mathbf{x}} = \alpha \mathbf{x}_\mathrm{r} + (1 - \alpha)\,\mathrm{sg}[\mathbf{x}_\mathrm{p}]$.
6: Query tracker field $\mathbf{v}_\mathrm{trk} = \boldsymbol{v}_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}, \mathbf{c})$.
7: Update $\boldsymbol{\theta}$ by descending $\|\mathbf{x}_\mathrm{p} - \mathrm{sg}[\mathbf{x}_\mathrm{p} + \mathbf{v}_\mathrm{trk}]\|_2^2$.
8: Update $\boldsymbol{\phi}$ by descending $\|\mathbf{v}_\mathrm{trk} - \mathrm{sg}[\mathbf{x}_\mathrm{r} - \mathrm{sg}[\mathbf{x}_\mathrm{p}]]\|_2^2$.
9: **until** convergence

data reference but retains exact within-batch generated-sample repulsion, making the fresh generated batch an operative optimization variable (Feng et al., 2026). Variational generative Wasserstein flows use JKO or dual formulations (Caucheteux et al., 2026), whereas Sinkhorn-Drifting and W-Flow construct minibatch cross- and self-transport plans (He et al., 2026; Han et al., 2026). W-Flow and Representation Distribution Matching are discussed here as related methods but are not numerical baselines in Tab. 2; we claim no numerical superiority over them. W-Flow is especially close at the training interface: it regresses a one-step generator toward detached particle updates obtained from minibatch Sinkhorn plans. At its exact endpoint, TBSM estimates each projectile's energy-distance field from one real and one independently generated source. A batch of B projectiles therefore uses $O(B)$ sampled source interactions and explicit frozen targets, without solving a minibatch transport problem; tracked scattering learns their conditional expectation online. Operationally, this sample-level interface requires one observed real source per condition-specific loss rather than a same-condition real minibatch, fitting text–image data with one image per caption. It does not remove the statistical limitation of observing one sample from each conditional law, discussed in Sec. G.

Guidance-free training interface. Classifier-free guidance is not intrinsic to the neighboring families, but leading implementations may build it into training: Drifting samples a guidance scale and unconditional real negatives, although its best reported FID uses the scale it identifies as "no CFG"; W-Flow adopts velocity guidance as its default; and text-to-image DMD distills teachers at fixed guidance scales (Deng et al., 2026; Han et al., 2026; Yin et al., 2024b; Ho & Salimans, 2022). The reported TBSM pipeline uses neither unconditional reference particles nor a guided teacher or sampler and introduces no separate guidance hyperparameter beyond its intrinsic (ρ, λ) controls. This is a claim about guidance-related pipeline simplicity, not that these methods cannot operate without CFG.

E Design-Map Interpretations

At the exact energy-distance endpoint, the population field is conservative. If a tracker constrained as $\boldsymbol{v}_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{c}) = \nabla_{\mathbf{x}} u_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{c})$ matches it throughout a connected domain, then $u_{\boldsymbol{\phi}}$ equals the negative first variation there up to an additive constant. Below, a modified displacement target gives a related potential interpretation along fake-to-real corridors. The displacement regression results below assume square-integrable endpoints and tracker predictions under the full sampling law, including the condition when it is averaged over, so their squared risks are finite.

E.1 Displacement Analogue of the GAN-Like State

Fake-to-real state at $\rho = 1, \lambda = 0$. At $\rho = 1, \lambda = 0$, Alg. 2 uses the unit-bearing target $\mathbf{b}_\mathrm{r}$ and queries the learned tracker along the fake-to-real corridor. Alg. 3 preserves this corridor but replaces the unit bearing with the displacement $\mathbf{x}_\mathrm{r} - \mathbf{x}_\mathrm{p}$, so it is not a parameter-only specialization. The resulting tracker learns both orientation and magnitude without a generated-source branch, while its detached output supervises the generator; feature normalization and the generator learning rate set the effective update scale.

Line-integral critic interpretation. The GAN-like connection becomes precise when the tracker is conservative, $\boldsymbol{v}_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{c}) = \nabla_{\mathbf{x}} u_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{c})$. Let $\Delta = \mathbf{x}_\mathrm{r} - \mathbf{x}_\mathrm{p}$ and $\mathbf{x}_\alpha = (1 - \alpha)\mathbf{x}_\mathrm{p} + \alpha \mathbf{x}_\mathrm{r}$. For the

potential identity, assume $u_{\boldsymbol{\phi}}(\cdot, \mathbf{c})$ is continuously differentiable, its endpoint values are integrable, and $\mathbb{E}\left[\|\nabla u_{\boldsymbol{\phi}}(\mathbf{x}_\alpha, \mathbf{c})\|_2^2\right] < \infty$ for independent $\alpha \sim \mathcal{U}(0,1)$. The line-integral identity along $\mathbf{x}_\alpha$ implies that, for a fixed generator,

$$\mathcal{L}_{\mathrm{trk}}^{\mathrm{disp}}(\boldsymbol{\phi}) = \mathbb{E}\left[\|\nabla_{\mathbf{x}} u_{\boldsymbol{\phi}}(\mathbf{x}_\alpha, \mathbf{c})\|_2^2\right] - 2\left(\mathbb{E}\left[u_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{r}}, \mathbf{c})\right] - \mathbb{E}\left[u_{\boldsymbol{\phi}}(\mathbf{x}_{\mathrm{p}}, \mathbf{c})\right]\right) + C, \tag{13}$$

where $C = \mathbb{E}\left[\|\Delta\|_2^2\right]$ is independent of $\boldsymbol{\phi}$ during the tracker update. Thus, for this conservative displacement analogue, tracker regression maximizes the real–generated potential gap minus one half of the corridor-averaged squared gradient norm; App. F.5.3 derives the identity and its conservative-field projection. This is still not identical to conventional GAN minimax training: the generator queries the field at a random corridor point rather than necessarily at $\mathbf{x}_{\mathrm{p}}$, and a general vector-valued tracker has no scalar-potential line-integral identity.

E.2 Reference Configurations and Matched Generator Updates

Here ρ mixes the instantaneous stochastic vector and learned tracker field, while λ jointly sets the intra-source coefficient in $\widehat{\mathbf{v}}_\lambda = \mathbf{b}_{\mathrm{r}} - \lambda \mathbf{b}_{\mathrm{s}}$ and the tracker-query range $\alpha \in [0, 1-\lambda]$. Because λ couples two changes, this is neither a factorized ablation nor a claim of exact equivalence to neighboring method families.

Instant scattering / Drift-like. At $\rho = 0, \lambda = 1$, the generator uses the instant energy-distance scattering update derived above; this is closest to Drift-like particle dynamics, but with a constant-size three-body estimator rather than a batch-level pairwise field.

Tracked scattering / energy-distance endpoint. At $\rho = 1, \lambda = 1$, the tracker is trained to estimate the same inter-minus-intra field; Cor. 1 gives its exact error decomposition.

Inter-only scattering and the denoising connection. At $\rho = 0, \lambda = 0$, the one-step update reduces to instantaneous attraction toward an independently sampled real source and is not itself a diffusion objective. An exact denoising-style reduction additionally requires paired $(\mathbf{x}_t, \mathbf{x}_0)$ across noise levels, conditioning on t, and a clean-sample, noise, or velocity target rather than a unit bearing; pure-noise input alone is insufficient.

Fake-to-real scattering and the GAN connection. At $\rho = 1, \lambda = 0$, Alg. 2 follows a learned field regressed from unit-bearing targets and queried along the fake-to-real corridor. The displacement analogue in Alg. 3 exposes the potential-tracker interpretation in (13), which does not transfer unchanged to the unit-bearing target.

Empirical comparison. At matched generator updates, $\rho = 1, \lambda = 1$ gives the lowest FID in Fig. 3, consistent with retaining attraction and self-repulsion while replacing source-sampling noise with a learned field. The $\rho = 0, \lambda = 0$ corner has the worst FID but highest IS, a pattern compatible with the contraction pressure induced by $\mathcal{F}_\lambda$ when $\lambda < 1$. The comparison is not factorized: $\rho = 0$ leaves the λ-specific field instantaneous, whereas changing λ also changes the tracker corridor. Interior ρ trades instant variance against tracker error, while λ near one retains self-repulsion; this motivates the near-endpoint benchmark settings in Sec. 4.3, without making the $\lambda < 1$ objective proper.

E.3 Connections to Related Distribution-Matching Objectives

The following relations require restrictions or input/target modifications beyond choosing (ρ, λ) and are interpretive rather than additional experimental algorithms.

A condition-only tracker and FD-loss. In the displacement analogue of Alg. 3, let $\mu_{\mathrm{r}}(\mathbf{c})$ and $\mu_{\boldsymbol{\theta}}(\mathbf{c})$ be the conditional real and generated feature means. Restricting the representation-space tracker to $\mathbf{c}$ alone gives the population minimizer $\boldsymbol{v}^\star(\mathbf{c}) = \mu_{\mathrm{r}}(\mathbf{c}) - \mu_{\boldsymbol{\theta}}(\mathbf{c})$. Assuming generator differentiability and valid differentiation under the expectation, this oracle makes the expected frozen-target generator gradient equal to the gradient of $\frac{1}{2}\|\mu_{\boldsymbol{\theta}}(\mathbf{c}) - \mu_{\mathrm{r}}(\mathbf{c})\|_2^2$, with the $1/2$ loss convention. This identity

concerns the expected gradient, not descent at every sampled update. This is an online, first-moment-only analogue of representation FD-loss (Yang et al., 2026), whose full Fréchet objective also matches feature covariances.

A displacement tracker and DMD. The displacement analogue in Alg. 3 resembles Distribution Matching Distillation (DMD) (Yin et al., 2024b) through an auxiliary model following the generated distribution and supplying a detached generator direction. DMD approximates a reverse-KL gradient from a fixed diffusion-teacher score and an online generated-distribution score at noise-perturbed generated samples; the analogue instead regresses fake-to-real displacements along straight corridors between independent endpoints, without teacher queries, score estimation, or a noise-time schedule.

Noisy-input inter-only scattering and perceptual supervision. If the inter-only corner receives a time-indexed noisy input $\mathbf{x}_t$, its paired clean sample $\mathbf{x}_0$, and condition t, its template approaches Perceptual Flow Matching (PFM) (Zhao et al., 2026) and PixelGen (Ma et al., 2026): each compares a clean prediction with $\mathbf{x}_0$ through frozen perceptual features. The objectives remain distinct because inter-only TBSM uses a normalized feature bearing and frozen-target regression, whereas PFM and PixelGen retain time-indexed flow/diffusion training with their perceptual losses.

F Derivation and Theoretical Foundations

Unless stated otherwise, distribution-level identities are for one fixed condition. They extend by averaging over $\mathbf{c}$ when the required expectations and interchanges are valid; condition-wise rates require corresponding integrable or uniform control. Parameter-space results apply to the aggregate conditional objective only when all their stated assumptions hold for it. Throughout this section, $\mathbf{x} \in \mathbb{R}^d$ denotes the vector supplied to the scattering objective. We assume finite first moments and set every bearing to zero at exact coincidence. Classical spatial and parameter gradients additionally require zero collision probability and valid differentiation under the expectation; with collision mass, zero is only a generalized-gradient selection. Population 2-Wasserstein flow statements further assume $P_t \in \mathcal{P}_2(\mathbb{R}^d)$ solves the continuity equation regularly enough for the 2-Wasserstein chain rule. At $\lambda = 1$, the code's denominator smoothing retains a proper objective; the frozen-target, tracker-error, and conditional convergence arguments apply to its corresponding smoothed field as detailed in Sec. F.5.1.

Our central question is what the tracker learns from sparse interactions and when this improves generator supervision. After deriving the exact field, we show that individual interactions retain noise at distributional agreement. Conditional-expectation regression extracts the predictable correction; at the full-corridor endpoint, its smoothed zero field identifies matching laws. A Gaussian location calculation then explains why weak local signal makes averaging valuable in high dimension. We distinguish these field-level properties from generator optimization, whose convergence results require the additional assumptions stated below.

F.1 Exact Scattering Field and Frozen-Target Equivalence

For one fixed condition, or in the unconditional case, write $P = P_{\boldsymbol{\theta}}$ and retain Q for the corresponding data law; the energy from the main text is

$$\mathcal{F}(P) = \mathbb{E}\left[\|\mathbf{x}_\mathrm{p} - \mathbf{x}_\mathrm{r}\|\right] - \frac{1}{2}\mathbb{E}\left[\|\mathbf{x}_\mathrm{p} - \mathbf{x}_\mathrm{s}\|\right] - \frac{1}{2}\mathbb{E}\left[\|\mathbf{x}_\mathrm{r} - \mathbf{x}'_\mathrm{r}\|\right], \tag{14}$$

where $\mathbf{x}_\mathrm{p}, \mathbf{x}_\mathrm{s} \sim P$ and $\mathbf{x}_\mathrm{r}, \mathbf{x}'_\mathrm{r} \sim Q$ are independent within and across laws. The real–real term is constant in P, while symmetry of the model–model term removes its factor $1/2$ under variation. Thus, up to an additive constant,

$$\frac{\delta \mathcal{F}}{\delta P}(\mathbf{x}) = \mathbb{E}\left[\|\mathbf{x} - \mathbf{x}_\mathrm{r}\|\right] - \mathbb{E}\left[\|\mathbf{x} - \mathbf{x}_\mathrm{s}\|\right], \tag{15}$$

$$\nabla_{\mathbf{x}} \frac{\delta \mathcal{F}}{\delta P}(\mathbf{x}) = \mathbb{E}\left[\frac{\mathbf{x} - \mathbf{x}_\mathrm{r}}{\|\mathbf{x} - \mathbf{x}_\mathrm{r}\|}\right] - \mathbb{E}\left[\frac{\mathbf{x} - \mathbf{x}_\mathrm{s}}{\|\mathbf{x} - \mathbf{x}_\mathrm{s}\|}\right], \tag{16}$$

$$\mathbf{v}(\mathbf{x}) = -\nabla_{\mathbf{x}} \frac{\delta \mathcal{F}}{\delta P}(\mathbf{x}) = \mathbb{E}\left[\frac{\mathbf{x}_\mathrm{r} - \mathbf{x}}{\|\mathbf{x}_\mathrm{r} - \mathbf{x}\|}\right] - \mathbb{E}\left[\frac{\mathbf{x}_\mathrm{s} - \mathbf{x}}{\|\mathbf{x}_\mathrm{s} - \mathbf{x}\|}\right]. \tag{17}$$

Evaluating the last line at a projectile and sampling one source from each law gives (4). For the aggregate conditional objective, differentiating through the generator yields

$$\nabla_{\boldsymbol{\theta}}\mathcal{F} = -\mathbb{E}_{\mathbf{c}\sim p_{\mathrm{data}}(\mathbf{c}),\,\mathbf{z}\sim p(\mathbf{z})}\left[J_{\boldsymbol{\theta}}\boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{c})^{\top}\mathbf{v}_{\boldsymbol{\theta}}(\boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{c}),\mathbf{c})\right]. \tag{18}$$

The symmetric model–model variation cancels the factor $1/2$ in $\mathcal{F}$, producing the same unit repulsion coefficient in both the population field and its generator pullback.

Frozen-target pullback. Fix $\mathbf{c}$ and current parameters $\bar{\boldsymbol{\theta}}$, abbreviate $Q = Q(\cdot \mid \mathbf{c})$, write $P_{\boldsymbol{\theta}} = (\boldsymbol{g}_{\boldsymbol{\theta}}(\cdot,\mathbf{c}))_{\#}p(\mathbf{z})$ and $\mathcal{F}_{\mathbf{c}}(\boldsymbol{\theta}) = \frac{1}{2}D_E^2(P_{\boldsymbol{\theta}},Q)$, and let $\mathcal{L}_{\mathrm{gen},\mathbf{c}}(\boldsymbol{\theta};\bar{\boldsymbol{\theta}})$ freeze all target-side quantities at $\bar{\boldsymbol{\theta}}$. For independent $\mathbf{x}_{\mathrm{r}} \sim Q$ and $\mathbf{z},\mathbf{z}_{\mathrm{s}} \sim p(\mathbf{z})$, set $\mathbf{x}_{\mathrm{p}} = \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\mathbf{z},\mathbf{c})$, $J_{\bar{\boldsymbol{\theta}}} = \nabla_{\boldsymbol{\theta}}\boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z},\mathbf{c})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}}$, $\mathbf{x}_{\mathrm{s}} = \boldsymbol{g}_{\bar{\boldsymbol{\theta}}}(\mathbf{z}_{\mathrm{s}},\mathbf{c})$, and $\widehat{\mathbf{v}} = \widehat{\mathbf{v}}_{\mathrm{scat}}(\mathbf{x}_{\mathrm{p}};\mathbf{x}_{\mathrm{r}},\mathbf{x}_{\mathrm{s}})$. We assume exact bearings, finite first moments, zero collision probability, almost-sure generator differentiability, and valid differentiation under the expectation, as in Thm. 1. For directly optimized particles, descending $\frac{1}{2}\left\|\mathbf{x}_{\mathrm{p}} - \mathrm{sg}[\mathbf{x}_{\mathrm{p}} + \widehat{\mathbf{v}}_{\mathrm{scat}}]\right\|_2^2$ takes the Euler step $\mathbf{x}_{\mathrm{p}} \leftarrow \mathbf{x}_{\mathrm{p}} + \eta\widehat{\mathbf{v}}_{\mathrm{scat}}$. For a generator, detach fixes this sampled field and the chain rule pulls it back through the projectile Jacobian.

Proof of Theorem 1. Independence of both sources from the projectile branch gives $\mathbb{E}\left[\widehat{\mathbf{v}} \mid \mathbf{x}_{\mathrm{p}},\mathbf{c}\right] = \mathbf{v}_{\bar{\boldsymbol{\theta}}}(\mathbf{x}_{\mathrm{p}},\mathbf{c})$ and, for the pullback, the same identity conditional on the projectile latent and its Jacobian; also $\|\widehat{\mathbf{v}}\|_2 \le 2$. Under the regularity assumptions in the theorem, all target-side quantities are fixed at $\bar{\boldsymbol{\theta}}$, so at the current parameters

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{gen},\mathbf{c}}(\boldsymbol{\theta};\bar{\boldsymbol{\theta}})\big|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}} = -\mathbb{E}\left[J_{\bar{\boldsymbol{\theta}}}^{\top}\widehat{\mathbf{v}}\right] = -\mathbb{E}\left[J_{\bar{\boldsymbol{\theta}}}^{\top}\mathbf{v}_{\bar{\boldsymbol{\theta}}}\right] = \nabla_{\boldsymbol{\theta}}\mathcal{F}_{\mathbf{c}}(\bar{\boldsymbol{\theta}}). \tag{19}$$

The middle equality uses conditional unbiasedness, and the last is (18). Without detach, derivatives through the generated source and target construction would add terms and change this identity. □

Dropping the conventional factor $1/2$ doubles the regression gradient and is absorbed by the learning rate.

F.2 Why Tracking Helps: Noise, Projection, and Identifiability

F.2.1 Nonvanishing interactions at distributional agreement

Tracking has a specific role even when the generated law already matches the data: the population correction vanishes, but a fresh scattering event need not become small. Fix $\lambda = 1$ and $P = Q$, and draw X_0, X_1, X_2 independently from Q. Write $B_{ij} = (X_j - X_i)/\left\|X_j - X_i\right\|_2$, with $B_{ij} = 0$ at coincidence, and set $Y = B_{01} - B_{02}$ and $V_Q = \mathbb{E}\left[\|Y\|_2^2\right]$. Let $a_2 = \Pr(X_0 = X_1)$ and $a_3 = \Pr(X_0 = X_1 = X_2)$ record possible atoms.

Proposition 2 (A noise floor at distributional agreement). *The exact scattering target satisfies* $\mathbb{E}\left[Y \mid X_0\right] = 0$*, but*

$$1 - a_2 \le V_Q \le \frac{4}{3} - 2a_2 + \frac{2}{3}a_3. \tag{20}$$

In particular, $1 \le V_Q \le 4/3$ *for every non-atomic* Q*, independently of dimension and scale;* $V_Q = 4/3$ *when such a law is supported on a line. These are bounds averaged over projectiles, not pointwise conditional-variance bounds.*

Proof. Conditional on X_0, the two source bearings are identically distributed and independent, giving the zero expectation. For a realized triple, define $C = \langle B_{01},B_{02}\rangle + \langle B_{10},B_{12}\rangle + \langle B_{20},B_{21}\rangle$. If the three points are distinct, C is the sum of the cosines of their triangle's angles, so $1 \le C \le 3/2$. Indeed, for angles $A_0 + A_1 + A_2 = \pi$, this sum is $1 + 4\prod_{i=0}^{2}\sin(A_i/2)$; concavity of $\log\sin$ bounds the product by $1/8$. A collinear triple has $C = 1$ by direct evaluation. Exactly two coincident points give $C = 1$, while three coincident points give $C = 0$. The probabilities of three distinct points and exactly two coincident points are $1 - 3a_2 + 2a_3$ and $3(a_2 - a_3)$, respectively; hence $1 - a_3 \le \mathbb{E}\left[C\right] \le \frac{3}{2}(1 - a_2)$. Exchangeability gives $\mathbb{E}\left[\langle B_{01},B_{02}\rangle\right] = \mathbb{E}\left[C\right]/3$, and $\mathbb{E}\left[\|B_{01}\|_2^2\right] = 1 - a_2$. Thus $V_Q = 2(1 - a_2) - \frac{2}{3}\mathbb{E}\left[C\right]$, proving the bounds. For a non-atomic law on a line, $C = 1$ almost surely, giving $V_Q = 4/3$. □

At this endpoint, the optimal population tracker is $h^\star = 0$, whereas its minimum raw regression risk $R(h) = \mathbb{E}\left[\|h(X_0) - Y\|_2^2\right]$ satisfies $R(h^\star) = V_Q$. For a fixed square-integrable tracker $h_{\boldsymbol{\phi}}$ and $G_\rho = (1-\rho)Y + \rho h_{\boldsymbol{\phi}}(X_0)$, conditional orthogonality gives

$$\mathbb{E}\left[\|G_\rho\|_2^2\right] = (1-\rho)^2 V_Q + \rho^2 \|h_{\boldsymbol{\phi}}\|_{L^2(Q)}^2 . \tag{21}$$

Averaging K conditionally independent source pairs at the same projectile leaves variance V_Q/K after averaging over projectiles; full oracle tracking instead gives a zero correction almost surely. Consequently, at agreement the oracle removes the scattering contribution to each frozen-target generator gradient, even though its regression loss generally remains positive. This does not assert that a finite online tracker attains the oracle or that optimizer momentum and weight decay vanish. The numerical bounds in (20) concern exact bearings: denominator smoothing retains the zero conditional expectation at $P = Q$ but changes the noise magnitude.

F.2.2 Conditional-expectation projection and finite tracker error

Let $\mathcal{H}$ denote the training history before a fresh triplet is drawn, and evaluate the current tracker before updating it on that triplet, so the current generator and tracker are fixed conditional on $\mathcal{H}$. Write $\mathbf{q} = (\tilde{\mathbf{x}}, \mathbf{c})$, $Y = \hat{\mathbf{v}}_\lambda$, $m_\lambda(\mathbf{q}) = \mathbb{E}[Y \mid \mathbf{q}, \mathcal{H}]$, and $h_{\boldsymbol{\phi}}(\mathbf{q}) = \boldsymbol{v}_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}, \mathbf{c})$. Assume $h_{\boldsymbol{\phi}}(\mathbf{q})$ is square integrable and define $G_\rho = (1-\rho)Y + \rho h_{\boldsymbol{\phi}}(\mathbf{q})$ for $\rho \in [0,1]$. The population tracker risk is $R(h) = \mathbb{E}\left[\|h(\mathbf{q}) - Y\|_2^2 \mid \mathcal{H}\right]$.

Theorem 2 (Tracked scattering as a conditional-expectation projection) . *Over all square-integrable functions measurable with respect to the query and history, the conditional expectation m_λ minimizes R, and the excess risk of the learned tracker is $R(h_{\boldsymbol{\phi}}) - R(m_\lambda) = \mathbb{E}\left[\|h_{\boldsymbol{\phi}}(\mathbf{q}) - m_\lambda(\mathbf{q})\|_2^2 \mid \mathcal{H}\right]$. Moreover, almost surely, the mixed target has the exact error decomposition*

$$\mathbb{E}\left[\|G_\rho - m_\lambda(\mathbf{q})\|_2^2 \mid \mathbf{q}, \mathcal{H}\right] = (1-\rho)^2 \operatorname{tr}\operatorname{Var}(Y \mid \mathbf{q}, \mathcal{H}) + \rho^2 \|h_{\boldsymbol{\phi}}(\mathbf{q}) - m_\lambda(\mathbf{q})\|_2^2 . \tag{22}$$

At $\lambda = 1$, $\mathbf{q} = (\mathbf{x}_\mathrm{p}, \mathbf{c})$ and $m_1(\mathbf{q}) = \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{x}_\mathrm{p}, \mathbf{c})$. In this case, the expected tracked velocity remains a strict energy-descent direction whenever $\rho \|h_{\boldsymbol{\phi}} - \mathbf{v}_{\boldsymbol{\theta}}\|_{L^2(P_{\boldsymbol{\theta}})} < \|\mathbf{v}_{\boldsymbol{\theta}}\|_{L^2(P_{\boldsymbol{\theta}})}$, conditionwise.

Proof. The conditional expectation is the L^2 orthogonal projection of Y onto the sigma-algebra generated by $\mathbf{q}$ together with $\mathcal{H}$. Hence, for every square-integrable $h(\mathbf{q})$, the cross term between $Y - m_\lambda(\mathbf{q})$ and $h(\mathbf{q}) - m_\lambda(\mathbf{q})$ vanishes conditional on $\mathcal{H}$. Expanding the tracker risk proves the minimizer and excess-risk statements.

For the mixed target, $G_\rho - m_\lambda(\mathbf{q}) = (1-\rho)(Y - m_\lambda(\mathbf{q})) + \rho(h_{\boldsymbol{\phi}}(\mathbf{q}) - m_\lambda(\mathbf{q}))$. The same conditional orthogonality removes the cross term, proving (22).

At $\lambda = 1$, the tracker query is the projectile, and Thm. 1 identifies m_1 with the exact population scattering vector field $\mathbf{v}_{\boldsymbol{\theta}}$. Suppress the condition and write $\boldsymbol{\epsilon}_{\boldsymbol{\phi}} = h_{\boldsymbol{\phi}} - \mathbf{v}_{\boldsymbol{\theta}}$. The conditional expectation of G_ρ is then $\mathbf{v}_{\boldsymbol{\theta}} + \rho\boldsymbol{\epsilon}_{\boldsymbol{\phi}}$. Whenever the Wasserstein chain rule is valid, the corresponding population flow satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(P_t) = -\|\mathbf{v}_t\|_{L^2(P_t)}^2 - \rho\langle \mathbf{v}_t, \boldsymbol{\epsilon}_{\boldsymbol{\phi},t}\rangle_{L^2(P_t)} \le -\|\mathbf{v}_t\|_{L^2(P_t)} \left(\|\mathbf{v}_t\|_{L^2(P_t)} - \rho\|\boldsymbol{\epsilon}_{\boldsymbol{\phi},t}\|_{L^2(P_t)}\right), \tag{23}$$

where the inequality is Cauchy–Schwarz. □

At $\lambda = 1$, full oracle tracking replaces Y by $m_1 = \mathbb{E}[Y \mid \mathbf{x}_\mathrm{p}, \mathbf{c}, \mathcal{H}]$, its Rao–Blackwellization with respect to $\sigma(\mathbf{x}_\mathrm{p}, \mathbf{c}, \mathcal{H})$; this preserves the energy field and removes source-sampling variance but not projectile-sampling variance.

For any $\lambda \in [0,1]$, define the projectile-conditioned source expectation $\overline{\mathbf{v}}_\lambda(\mathbf{x}_\mathrm{p}, \mathbf{c}) = \mathbb{E}[Y \mid \mathbf{x}_\mathrm{p}, \mathbf{c}, \mathcal{H}]$, which is the negative Wasserstein-gradient field of (46) before tracker conditioning. Then

$$\mathbb{E}[G_\rho \mid \mathbf{x}_\mathrm{p}, \mathbf{c}, \mathcal{H}] = (1-\rho)\overline{\mathbf{v}}_\lambda(\mathbf{x}_\mathrm{p}, \mathbf{c}) + \rho\mathbb{E}[h_{\boldsymbol{\phi}}(\mathbf{q}) \mid \mathbf{x}_\mathrm{p}, \mathbf{c}, \mathcal{H}] . \tag{24}$$

For $\lambda < 1$, the query- and projectile-generated sigma-algebras are generally not nested, so even for an oracle tracker $h_\phi = m_\lambda$, the second conditional expectation in (24) need not equal $\overline{\mathbf{v}}_\lambda$. At $\lambda = 1$, $\mathbf{q} = (\mathbf{x}_\mathrm{p}, \mathbf{c})$, so the two expectations coincide for an oracle tracker. Thus (22) remains exact around m_λ for every λ, but for $\lambda < 1$ it does not imply descent of either the energy distance or $\mathcal{F}_\lambda$ after generator pullback.

If mixing can depend on the query, (22) is minimized pointwise by $\rho^\star(\mathbf{q}) = s_\lambda(\mathbf{q})/(s_\lambda(\mathbf{q}) + e_\phi(\mathbf{q}))$ whenever the denominator is positive, where $s_\lambda(\mathbf{q}) = \operatorname{tr}\operatorname{Var}(Y \mid \mathbf{q}, \mathcal{H})$ and $e_\phi(\mathbf{q}) = \|h_\phi(\mathbf{q}) - m_\lambda(\mathbf{q})\|_2^2$. For the query-independent weight used in the algorithm, let $S_\lambda = \mathbb{E}[s_\lambda(\mathbf{q}) \mid \mathcal{H}]$ and $E_\phi = \mathbb{E}[e_\phi(\mathbf{q}) \mid \mathcal{H}]$. When $S_\lambda + E_\phi > 0$, integrating (22) and minimizing its quadratic in ρ gives $\rho^\star_{\mathrm{const}} = S_\lambda/(S_\lambda + E_\phi)$ and minimum error $S_\lambda E_\phi/(S_\lambda + E_\phi)$. Thus full tracking ($\rho = 1$) improves on the instant target, relative to m_λ, precisely when its excess regression risk E_ϕ is below the residual target variance S_λ. At $\lambda = 1$, this residual is exactly the source-sampling noise around the energy-distance field. At this endpoint, (22) specializes to Cor. 1; in the oracle case, $E_\phi = 0$ and $\rho^\star_{\mathrm{const}} = 1$.

A two-draw tracker diagnostic. Fix $\lambda = 1$ and condition on a history $\mathcal{H}$. For each projectile query $\mathbf{q} = (\mathbf{x}_\mathrm{p}, \mathbf{c})$, hold the generator, tracker, query, and $h_\phi(\mathbf{q})$ fixed while independently resampling two complete source pairs, producing conditionally iid targets Y_1, Y_2. Then

$$S_1 = \frac{1}{2}\mathbb{E}\left[\|Y_1 - Y_2\|_2^2 \mid \mathcal{H}\right], \qquad E_\phi = \mathbb{E}\left[\langle h_\phi(\mathbf{q}) - Y_1, h_\phi(\mathbf{q}) - Y_2\rangle \mid \mathcal{H}\right]. \tag{25}$$

Indeed, conditional on $(\mathbf{q}, \mathcal{H})$, writing $Y_i = m_1(\mathbf{q}) + \xi_i$ with independent centered residuals gives the two identities after expansion and averaging over $\mathbf{q}$. Consequently, when $S_1 + E_\phi > 0$, the field-MSE-optimal constant mixture is $\rho^\star_{\mathrm{const}} = S_1/(S_1 + E_\phi)$, and full tracking has lower field MSE than the instant target exactly when $E_\phi < S_1$. Neither identity requires the unknown field m_1, so when the required conditional redraws are available, paired Monte Carlo averages provide unbiased snapshot estimates of S_1 and E_ϕ. The resulting ratio is only a plug-in diagnostic (and may require clipping at finite sample size), not an unbiased estimate of $\rho^\star_{\mathrm{const}}$ or a certificate of generator convergence or sample quality.

At $\lambda = 1$, the same orthogonal decomposition holds after the field is pulled back through the generator. Let $J = \nabla_{\boldsymbol{\theta}} \boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{c})$ and $\mathbf{v}(\mathbf{q}) = m_1(\mathbf{q})$. Assume $J^\top(Y - \mathbf{v}(\mathbf{q}))$ and $J^\top(h_\phi(\mathbf{q}) - \mathbf{v}(\mathbf{q}))$ are square integrable, and define, conditional on $\mathcal{H}$, $S_J = \mathbb{E}\left[\left\|J^\top(Y - \mathbf{v}(\mathbf{q}))\right\|_2^2 \mid \mathcal{H}\right]$ and $E_{J,\phi} = \mathbb{E}\left[\left\|J^\top(h_\phi(\mathbf{q}) - \mathbf{v}(\mathbf{q}))\right\|_2^2 \mid \mathcal{H}\right]$.

Corollary 2 (Generator-update tracker decomposition). *If $\mathbb{E}[Y - \mathbf{v}(\mathbf{q}) \mid \mathbf{q}, J, \mathcal{H}] = 0$, as ensured by the independent projectile and source draws in Alg. 2, then*

$$\mathbb{E}\left[\left\|J^\top(G_\rho - \mathbf{v}(\mathbf{q}))\right\|_2^2 \mid \mathcal{H}\right] = (1-\rho)^2 S_J + \rho^2 E_{J,\phi}. \tag{26}$$

Proof. Conditional on $(\mathbf{q}, J, \mathcal{H})$, both J and $h_\phi(\mathbf{q}) - \mathbf{v}(\mathbf{q})$ are fixed. The assumed conditional-expectation identity makes the cross term vanish after multiplication by $J^\top$, proving (26). □

Thus, when $S_J + E_{J,\phi} > 0$, the query-independent mixture minimizing the per-sample Jacobian-pullback error is $\rho^\star_J = S_J/(S_J + E_{J,\phi})$; full tracking improves on the instant update exactly when $E_{J,\phi} < S_J$. This parameter-space criterion can differ from the output-space criterion because the generator Jacobian suppresses or amplifies different field directions.

F.2.3 Identifiability of the full-corridor tracker field

For the full-corridor endpoint $\lambda = 0$, the tracker observes Z and the condition, but not the interpolation coefficient A. Fix one condition, take independent $X \sim P$, $R \sim Q$, and $A \sim \mathcal{U}[0,1]$, and define

$$Z = (1-A)X + AR, \qquad T_\varepsilon = \frac{R - X}{\|R - X\|_2 + \varepsilon}, \qquad m_\varepsilon(Z) = \mathbb{E}[T_\varepsilon \mid Z], \quad \varepsilon > 0. \tag{27}$$

The target is bounded, so this result does not require the moment assumptions used for the energy-distance analysis. At agreement, exchanging (X, R, A) with $(R, X, 1-A)$ leaves the query fixed and

reverses the target. This cancellation also holds for exact bearings and square-integrable displacement targets. For example, for exact bearings and $P = Q$ uniform on $[-1, 1]$, the instantaneous attraction-only field is $-x$, whereas the oracle corridor field is zero. Prop. 1 proves the converse for the smoothed bearing: cancellation throughout the corridor forces the endpoint laws to agree.

Zero-Flow Encoders establishes a zero-field criterion for independent-coupling midpoint displacement fields, and Midpoint Generative Models develops related discrepancies for one-step generation (Wang et al., 2026; Shlenskii et al., 2026). Our statement concerns the uniformly sampled full corridor with a denominator-smoothed bearing. Its proof accounts for the distance-dependent endpoint reweighting introduced by normalization.

Applying Prop. 1 separately at each condition, if $m_\varepsilon(Z, \mathbf{c}) = 0$ for almost every joint corridor query and condition and the tracker does not observe A separately, then $P(\cdot \mid \mathbf{c}) = Q(\cdot \mid \mathbf{c})$ for $p_{\mathrm{data}}(\mathbf{c})$-almost every $\mathbf{c}$.

Proof of Proposition 1. If $P = Q$, swapping (X, R, A) with $(R, X, 1 - A)$ preserves Z and negates T_ε. Hence $m_\varepsilon(Z) = 0$ almost surely.

Conversely, suppose $m_\varepsilon(Z) = 0$ almost surely. For every $\varphi \in C_c^\infty(\mathbb{R}^d)$, conditional expectation and the fundamental theorem of calculus along the segment from X to R give

$$0 = \mathbb{E}\left[T_\varepsilon^\top \nabla\varphi(Z)\right] = \mathbb{E}\left[\frac{\varphi(R) - \varphi(X)}{\|R - X\|_2 + \varepsilon}\right]. \tag{28}$$

Define $w(x, y) = (\|x - y\|_2 + \varepsilon)^{-1}$ and $K_\mu(x) = \int w(x, y)\mu(\mathrm{d}y)$. Equation (28) says that the finite measures $K_P(x)Q(\mathrm{d}x)$ and $K_Q(x)P(\mathrm{d}x)$ are equal. Since $0 < K_\mu(x) \le 1/\varepsilon$, these measures imply P and Q are equivalent. Let $f = \mathrm{d}P/\mathrm{d}Q$; their equality yields $K_P(x) = f(x)K_Q(x)$ for Q-almost every x.

The transition kernel $\Pi(x, \mathrm{d}y) = w(x, y)Q(\mathrm{d}y)/K_Q(x)$ then satisfies $\Pi f = f$. It is reversible with respect to the finite measure $\pi(\mathrm{d}x) = K_Q(x)Q(\mathrm{d}x)$ because $\pi(\mathrm{d}x)\Pi(x, \mathrm{d}y) = w(x, y)Q(\mathrm{d}x)Q(\mathrm{d}y)$ is symmetric. Since $K_Q \le 1/\varepsilon$ and $\int f\,\mathrm{d}Q = 1$, f is π-integrable, and $\psi(u) = (1+u)^{-1}$ is bounded and strictly convex on $[0, \infty)$. Jensen's inequality and invariance give

$$\int \psi(f)\,\mathrm{d}\pi = \int \psi(\Pi f)\,\mathrm{d}\pi \le \int \Pi\psi(f)\,\mathrm{d}\pi = \int \psi(f)\,\mathrm{d}\pi.$$

Equality in strict Jensen implies that f is constant under $\Pi(x, \cdot)$ for π-almost every x. As $w(x, y) > 0$, these transition laws are equivalent to Q, so f is constant Q-almost everywhere. Since $\int f\,\mathrm{d}Q = 1$, the constant is one and $P = Q$. □

The field therefore contains enough information to distinguish any unequal endpoint laws at this full-corridor endpoint, although each displacement is normalized before conditional averaging. This is a qualitative property of the population field: it does not bound a distributional discrepancy by a small field norm, characterize partially sampled corridors, or establish the $\varepsilon = 0$ converse. Nor does it by itself guarantee a descending generator update. For example, let $Q = (\delta_{-1} + \delta_1)/2$ and $g_\theta(\xi) = \theta\xi$ for an independent equiprobable $\xi \in \{-1, 1\}$. At $\theta = 0$, $P_0 = \delta_0 \ne Q$, but $m_\varepsilon(z) = \mathrm{sign}(z)/(1 + \varepsilon)$ for corridor points $z \ne 0$. In the fully tracked $\lambda = 0$ update, the query is $Z = AR$, independent of ξ, and, with the $1/2$ loss convention, the expected frozen-target parameter gradient is

$$-\mathbb{E}\left[\xi\, m_\varepsilon(AR)\right] = 0. \tag{29}$$

The feature-space field is nonzero, yet the generator Jacobian cancels it in expectation. Thus field identifiability and parameter-space optimization are distinct properties.

F.2.4 Local signal-to-noise in a Gaussian location model

We now connect the noise floor to the difficulty of estimating small corrections in high dimension. Fix $\lambda = 1$ with exact bearings, let $Q_d = \mathcal{N}(0, I_d)$ and $P_a = \mathcal{N}(a, I_d)$, and use the location generator $g_a(\xi) = a + \xi$. For independent $\Xi_0, \Xi_s, R \sim \mathcal{N}(0, I_d)$, set $X_0 = a + \Xi_0$, $X_s = a + \Xi_s$, and $Y_a = (R - X_0)/\|R - X_0\|_2 - (X_s - X_0)/\|X_s - X_0\|_2$. Thus $-Y_a$ is an unbiased one-triplet estimator of $\nabla_a\mathcal{F}(P_a)$ under the $1/2$ loss convention. Define its signal-to-RMS-noise ratio by $\mathrm{SNR}_d(a) = \|\nabla_a\mathcal{F}(P_a)\|_2 / \sqrt{\mathrm{tr}\,\mathrm{Var}(-Y_a)}$.

Corollary 3 (Weak local signal despite bounded triplet noise) . *For each $d \geq 2$, as $a \to 0$,*

$$\nabla_a \mathcal{F}(P_a) = c_d a + o(\|a\|_2), \qquad c_d = \frac{d-1}{2d}\frac{\Gamma((d-1)/2)}{\Gamma(d/2)} \sim \frac{1}{\sqrt{2d}}. \tag{30}$$

Writing $V_{Q_d} = \mathbb{E}\left[\|Y_0\|_2^2\right] \in [1, 4/3]$, every unit vector u satisfies

$$\lim_{r \downarrow 0} \frac{\mathrm{SNR}_d(ru)}{r} = \frac{c_d}{\sqrt{V_{Q_d}}} = \Theta(d^{-1/2}). \tag{31}$$

The local limit is taken at fixed d, followed by the dimension asymptotic.

Proof. Write $W = R - \Xi_0 \sim \mathcal{N}(0, 2I_d)$. The expected model–model bearing vanishes by exchange symmetry, so the location gradient is $g(a) = \mathbb{E}\left[(a - W)/\|a - W\|_2\right]$. To justify differentiation, view this expectation as the convolution of the bounded bearing map with a Gaussian density. Differentiate the density, then integrate by parts; the bearing derivative is locally integrable for $d \geq 2$, and the boundary term at a ball of radius δ around zero vanishes as $O(\delta^{d-1})$. Isotropy then gives

$$Dg(0) = \mathbb{E}\left[\frac{I_d}{\|W\|_2} - \frac{WW^\top}{\|W\|_2^3}\right] = \frac{d-1}{d}\mathbb{E}\left[\|W\|_2^{-1}\right] I_d.$$

Since $\mathbb{E}\left[\|W\|_2^{-1}\right] = \frac{1}{2}\Gamma((d-1)/2)/\Gamma(d/2)$, (30) follows; the asymptotic uses the gamma-ratio expansion. Coupling the same Gaussian draws for all a yields $Y_a \to Y_0$ almost surely, with $\|Y_a\|_2 \leq 2$. Bounded convergence and Prop. 2 give $\operatorname{tr}\operatorname{Var}(-Y_a) \to V_{Q_d} \in [1, 4/3]$, proving (31). □

The local signal shrinks linearly with the mean error, with sensitivity of order $d^{-1/2}$, while the instantaneous noise approaches a nonzero constant. A dimension-free variance bound therefore does not imply dimension-free signal-to-noise. At this endpoint, oracle tracking preserves the population gradient and removes source-sampling noise; a learned tracker improves supervision when its excess risk is below that noise, as quantified above. This explains the value of learning the conditional field near agreement without asserting that an online neural tracker automatically attains the required accuracy. The calculation concerns exact bearings in a Gaussian location family and gives no uniform joint (a, d) limit or quantitative guarantee for image features.

F.3 Population Dissipation and Idealized Optimization Guarantees

F.3.1 Exact-flow dissipation

Let P_t solve $\partial_t P_t + \nabla \cdot (P_t \mathbf{v}_t) = 0$ with $\mathbf{v}_t = -\nabla(\delta\mathcal{F}/\delta P_t)$.

Whenever the 2-Wasserstein chain rule is valid,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(P_t) = -\int \|\mathbf{v}_t(\mathbf{x})\|_2^2 \, P_t(\mathrm{d}\mathbf{x}) \leq 0. \tag{32}$$

Indeed, writing $\phi_t = \delta\mathcal{F}/\delta P_t$, the chain rule, continuity equation, and integration by parts give $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(P_t) = \int \nabla\phi_t^\top \mathbf{v}_t \, \mathrm{d}P_t = -\int \|\mathbf{v}_t\|_2^2 \, \mathrm{d}P_t$. Consequently, for every $T > 0$,

$$\inf_{0 \leq t \leq T} \int \|\mathbf{v}_t\|_2^2 \, \mathrm{d}P_t \leq \frac{1}{T}\int_0^T \int \|\mathbf{v}_t\|_2^2 \, \mathrm{d}P_t \, \mathrm{d}t \leq \frac{\mathcal{F}(P_0)}{T}. \tag{33}$$

This is field stationarity, not distributional or finite-generator convergence. Properness by itself is insufficient to rule out nonoptimal stationary measures, so the slope condition used below is substantive. A global slope inequality is not automatic because kernel discrepancies are generally not geodesically convex in Wasserstein geometry. Related Riesz-flow theory proves global exponential convergence in the Coulomb regime under regularity assumptions (Boufadène & Vialard, 2025) and local polynomial guarantees in higher-order Sobolev regimes under stronger assumptions (Chizat et al., 2026); neither directly supplies the energy-distance slope inequality used next.

F.3.2 TRACKED POPULATION FLOW

We now close the population-level chain for the exact energy-distance endpoint $\lambda = 1$. Let Q be fixed, write $\mathcal{F}(P) = \frac{1}{2}D_E^2(P, Q)$, and let $\mathbf{v}_t = -\nabla(\delta\mathcal{F}/\delta P_t)$ be its exact scattering field. For a fixed $\rho \in [0, 1]$, let h_t be the current population tracker, set $e_t = h_t - \mathbf{v}_t$, and suppose P_t solves the continuity equation with the mean mixed velocity $u_t = (1-\rho)\mathbf{v}_t + \rho h_t = \mathbf{v}_t + \rho e_t$. Let R_t denote the population tracker risk from Sec. F.2.2 at P_t. The conditional-expectation projection gives $R_t(h_t) - R_t(\mathbf{v}_t) = \|e_t\|^2_{L^2(P_t)}$. In addition to the chain-rule regularity assumed above, suppose that along this trajectory there are constants $\gamma \in [0, 1)$, $c > 0$, and $q \geq 1$ such that

$$\rho\sqrt{R_t(h_t) - R_t(\mathbf{v}_t)} \leq \gamma \|\mathbf{v}_t\|_{L^2(P_t)}, \qquad \|\mathbf{v}_t\|^2_{L^2(P_t)} \geq c\,\mathcal{F}(P_t)^q \quad \text{for all } t \geq 0. \tag{34}$$

The first inequality directly controls tracker excess risk relative to field energy; the second is a trajectory-wise slope inequality, with exponential and polynomial regimes at $q = 1$ and $q > 1$, respectively.

Theorem 3 (Tracked population-flow convergence from excess risk). *Under the preceding assumptions, $\mathcal{F}(P_t) \to 0$ and hence $D_E(P_t, Q) \to 0$. If $\mathcal{F}(P_0) > 0$, the quantitative bound is*

$$\mathcal{F}(P_t) \leq \begin{cases} \mathcal{F}(P_0)\exp[-c(1-\gamma)t], & q = 1, \\ \left[\mathcal{F}(P_0)^{1-q} + c(q-1)(1-\gamma)t\right]^{-1/(q-1)}, & q > 1. \end{cases} \tag{35}$$

Proof. Since $\mathbf{v}_t = -\nabla(\delta\mathcal{F}/\delta P_t)$, the Wasserstein chain rule and Cauchy–Schwarz give

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(P_t) = -\langle \mathbf{v}_t, u_t\rangle_{L^2(P_t)} \leq -(1-\gamma)\|\mathbf{v}_t\|^2_{L^2(P_t)} \leq -c(1-\gamma)\mathcal{F}(P_t)^q. \tag{36}$$

If $\mathcal{F}(P_0) > 0$, integrating this scalar differential inequality proves (35). If $\mathcal{F}(P_0) = 0$, the same inequality and nonnegativity imply $\mathcal{F}(P_t) = 0$ for all $t \geq 0$. In either case, properness and $D_E^2(P_t, Q) = 2\mathcal{F}(P_t)$ give the stated convergence. □

Broader quantile-space analyses characterize negative-distance MMD flows on the real line (Duong et al., 2026); the following elementary bound isolates a sufficient slope regime for tracked convergence.

Proposition 3 (A concrete one-dimensional slope regime). *Let P_t and Q be supported on a common finite interval $I \subset \mathbb{R}$, and suppose P_t has a density $p_t \geq m > 0$ almost everywhere on I, uniformly in t. Writing F_t and F_Q for their cumulative distribution functions, the exact energy-distance field satisfies*

$$\|\mathbf{v}_t\|^2_{L^2(P_t)} = 4\int_I (F_t - F_Q)^2 p_t\,\mathrm{d}x \geq 4m\,\mathcal{F}(P_t). \tag{37}$$

Hence the slope condition in Thm. 3 holds with $q = 1$ and $c = 4m$; under the theorem's remaining assumptions, $\mathcal{F}(P_t) \leq \mathcal{F}(P_0)\exp[-4m(1-\gamma)t]$.

Proof. The one-dimensional energy-distance identity gives $D_E^2(P_t, Q) = 2\int_I (F_t - F_Q)^2\,\mathrm{d}x$, and therefore $\mathcal{F}(P_t) = \int_I (F_t - F_Q)^2\,\mathrm{d}x$. Differentiating the first variation in (15) at Lebesgue-almost every x yields $\mathbf{v}_t(x) = 2(F_t(x) - F_Q(x))$. Integrating its square against p_t and using $p_t \geq m$ proves (37); the rate then follows from Thm. 3. □

Meaning for tracker training. The first condition in (34) is a population-level certificate stated in the tracker's own regression geometry: full tracking requires its root excess risk to remain below the field norm, while mixing by $\rho < 1$ attenuates tracker error. The algorithm alone does not guarantee that a finite neural tracker maintains this relative certificate as the signal vanishes. Absolute-error neighborhoods, smoothing, reweighting, and corridor conditioning are unified by Prop. 8 below.

F.3.3 FINITE-GENERATOR REFERENCE RECURSION

The results below analyze the stated first-order stochastic recursion and isolate how field-estimation error enters. They do not cover AdamW momentum or preconditioning or the $\lambda < 1$ corridor dynamics, and they do not establish the displayed bias and variance conditions for EMA-evolving quantities or the jointly trained tracker.

Suppose the parameterized objective $\mathcal{F}(\boldsymbol{\theta})$ is lower bounded by $\mathcal{F}_\star$ and L-smooth, with $L > 0$, on a region containing every update segment between $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{t+1}$. Take deterministic initial parameters with finite $\mathcal{F}(\boldsymbol{\theta}_0)$, and assume the parameter gradients and biases below are square integrable at each finite iteration. Let $\mathcal{H}_t$ be the history before iteration t, and absorb any fixed algorithmic loss scaling into the learning rate. Write the parameter update as $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t g_t$ and assume $\mathbb{E}[g_t \mid \mathcal{H}_t] = \nabla\mathcal{F}(\boldsymbol{\theta}_t) + b_t$ and $\mathbb{E}\left[\|g_t - \mathbb{E}[g_t \mid \mathcal{H}_t]\|_2^2 \mid \mathcal{H}_t\right] \le \sigma^2$. Here b_t collects systematic field-estimation error, while σ^2 controls the residual stochastic variation.

Proposition 4 (Optimization with finite field-estimation error) . *For constant step size $0 < \eta_t = \eta \le 1/L$, the stochastic update satisfies*

$$\min_{0\le t<T} \mathbb{E}\left[\|\nabla_{\boldsymbol{\theta}}\mathcal{F}(\boldsymbol{\theta}_t)\|_2^2\right] \le \frac{2(\mathcal{F}(\boldsymbol{\theta}_0) - \mathcal{F}_\star)}{\eta T} + L\eta\sigma^2 + \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|b_t\|_2^2\right]. \tag{38}$$

Proof. Let $a_t = \nabla\mathcal{F}(\boldsymbol{\theta}_t)$, $\bar{g}_t = \mathbb{E}[g_t \mid \mathcal{H}_t] = a_t + b_t$, and let $\mathbb{E}_t$ condition on $\mathcal{H}_t$. By L-smoothness,

$$\begin{aligned}\mathbb{E}_t[\mathcal{F}(\boldsymbol{\theta}_{t+1})] &\le \mathcal{F}(\boldsymbol{\theta}_t) - \eta\langle a_t, \bar{g}_t\rangle + \frac{L\eta^2}{2}\mathbb{E}_t[\|g_t\|_2^2] \\ &\le \mathcal{F}(\boldsymbol{\theta}_t) - \frac{\eta}{2}\|a_t\|_2^2 + \frac{\eta}{2}\|b_t\|_2^2 + \frac{L\eta^2\sigma^2}{2},\end{aligned} \tag{39}$$

where the second inequality uses $-\langle a_t, \bar{g}_t\rangle = (\|b_t\|_2^2 - \|a_t\|_2^2 - \|\bar{g}_t\|_2^2)/2$, $\mathbb{E}_t\|g_t\|_2^2 \le \|\bar{g}_t\|_2^2 + \sigma^2$, and $\eta \le 1/L$. Taking expectations, summing over $t = 0, \dots, T-1$, and using $\mathcal{F}(\boldsymbol{\theta}_T) \ge \mathcal{F}_\star$ proves (38). □

Under the local-gradient regularity conditions of Thm. 1, the instant update at $(\rho, \lambda) = (0, 1)$ has $b_t = 0$: relative to $\mathcal{F}$ for exact bearings and to $\mathcal{F}_\varepsilon$ for smoothed bearings. For the exact-bearing tracked algorithm at $\lambda = 1$, let $e_t(\mathbf{x}, \mathbf{c}) = h_{\phi,t}(\mathbf{x}, \mathbf{c}) - \mathbf{v}_t(\mathbf{x}, \mathbf{c})$. The conditional expectation is $\mathbf{v}_t + \rho e_t$, so the parameter-gradient bias in Prop. 4 is $b_t = -\rho\mathbb{E}\left[J_t^\top e_t \mid \mathcal{H}_t\right]$, where $J_t = \nabla_{\boldsymbol{\theta}}\boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_t, \mathbf{c}_t)$. Jensen's inequality gives $\|b_t\|_2^2 \le \rho^2\mathbb{E}\left[\left\|J_t^\top e_t\right\|_2^2 \mid \mathcal{H}_t\right]$. Combining this estimate with (38) makes the Jacobian-weighted dependence explicit:

$$\min_{0\le t<T} \mathbb{E}\left[\|\nabla_{\boldsymbol{\theta}}\mathcal{F}(\boldsymbol{\theta}_t)\|_2^2\right] \le \frac{2(\mathcal{F}(\boldsymbol{\theta}_0) - \mathcal{F}_\star)}{\eta T} + L\eta\sigma^2 + \frac{\rho^2}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|J_t^\top e_t\right\|_2^2\right]. \tag{40}$$

If $\|J_t\|_{\mathrm{op}} \le \kappa_J$ almost surely, the final term is further bounded by $\rho^2\kappa_J^2 T^{-1}\sum_t \mathbb{E}\left[\|e_t\|_2^2\right]$. Thus the Jacobian-weighted tracker error in Cor. 2 controls the optimization neighborhood; under the operator-norm bound, the output-space excess risk in Thm. 2 does as well. An oracle tracker eliminates this term, but residual stochastic variance σ^2 may remain. At $\lambda \ne 1$, intra-source reweighting changes the reference objective from $\mathcal{F}$ to $\mathcal{F}_\lambda$, and query conditioning can change the mean direction even relative to $\mathcal{F}_\lambda$. Choosing $\mathcal{F}_\varepsilon$ as the reference removes smoothing bias at $\lambda = 1$; relative to the unsmoothed $\mathcal{F}$, the field difference enters through b_t when bounded.

Under explicit additional assumptions, this reference recursion also yields an expected convergence result. Consider the exact-bearing $\lambda = 1$ objective, assume the target is realizable, i.e., some $\boldsymbol{\theta}^\star$ satisfies $\mathcal{F}(\boldsymbol{\theta}^\star) = 0$, and suppose the parameterized objective satisfies $\|\nabla\mathcal{F}(\boldsymbol{\theta})\|_2^2 \ge 2\mu\mathcal{F}(\boldsymbol{\theta})$ with $\mu > 0$ throughout the training region. Assume also that $\|b_t\|_2 \le \chi\|\nabla\mathcal{F}(\boldsymbol{\theta}_t)\|_2$ almost surely for some $\chi \in [0, 1)$. For tracked scattering, the stronger but directly interpretable condition $\rho^2\mathbb{E}\left[\left\|J_t^\top e_t\right\|_2^2 \mid \mathcal{H}_t\right] \le \chi^2\|\nabla\mathcal{F}(\boldsymbol{\theta}_t)\|_2^2$ is sufficient by Jensen's inequality. Let positive deterministic step sizes satisfy $\eta_t \le \min\{(1-\chi)/(L(1+\chi)^2), [\mu(1-\chi)]^{-1}\}$.

Corollary 4 (Finite-generator convergence under PL and relative tracking) . *Under the preceding assumptions, if $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$, then*

$$\mathbb{E}\left[\mathcal{F}(\boldsymbol{\theta}_{t+1}) \mid \mathcal{H}_t\right] \le \left[1 - \mu(1-\chi)\eta_t\right] \mathcal{F}(\boldsymbol{\theta}_t) + \frac{L\eta_t^2\sigma^2}{2}. \tag{41}$$

Consequently, $\mathbb{E}\left[D_E^2(P_{\boldsymbol{\theta}_t}, Q)\right] \to 0$.

Proof. Write $a_t = \nabla\mathcal{F}(\boldsymbol{\theta}_t)$. The relative-bias condition gives $\langle a_t, a_t + b_t\rangle \ge (1-\chi)\left\|a_t\right\|_2^2$ and $\left\|a_t + b_t\right\|_2^2 \le (1+\chi)^2\left\|a_t\right\|_2^2$. Applying L-smoothness as in the proof of Prop. 4 and using the step-size bound therefore yields

$$\mathbb{E}\left[\mathcal{F}(\boldsymbol{\theta}_{t+1}) \mid \mathcal{H}_t\right] \le \mathcal{F}(\boldsymbol{\theta}_t) - \frac{(1-\chi)\eta_t}{2}\left\|a_t\right\|_2^2 + \frac{L\eta_t^2\sigma^2}{2}. \tag{42}$$

The parameter-space PL inequality proves (41). After taking expectations, set $x_t = \mathbb{E}\left[\mathcal{F}(\boldsymbol{\theta}_t)\right]$, $a = \mu(1-\chi)$, and $d = L\sigma^2/2$. Then $x_t + d\sum_{k=t}^{\infty}\eta_k^2$ decreases by at least $a\eta_t x_t$; convergence of this sequence, the vanishing tail, and $\sum_t \eta_t = \infty$ force $x_t \to 0$. Finally, $D_E^2(P_{\boldsymbol{\theta}_t}, Q) = 2\mathcal{F}(\boldsymbol{\theta}_t)$. □

The same proof applies to $\mathcal{F}_\varepsilon$ under the corresponding smoothness, realizability, PL, and relative-tracking assumptions, yielding $\mathbb{E}\left[\mathcal{F}_\varepsilon(\boldsymbol{\theta}_t)\right] \to 0$.

Interpretation. The unconstrained exact and smoothed $\lambda = 1$ energies have the unique optimum $P = Q$; Thm. 3 and its smoothed counterpart show that sufficiently accurate tracking preserves convergence whenever the corresponding flow satisfies the stated slope condition. A finite generator instead restricts the reachable distributions to $\{(\boldsymbol{g}_{\boldsymbol{\theta}})_{\#}p(\mathbf{z})\}$, and SGD optimizes a nonconvex parameterization; finite tracking, corridor conditioning, intra-source reweighting, and the choice of comparison geometry separate deployment from the exact full-space flow. Accordingly, within this reference recursion, (38) gives a finite-error stationarity bound, while Cor. 4 obtains expected energy-distance convergence only after adding realizability, parameter-space PL geometry, relative tracker accuracy, and diminishing step sizes.

F.4 Source-Law Necessity and Variance Optimality

Fix a projectile $\mathbf{x}_{\mathrm{p}}$ and let $\beta_{\mathbf{x}_{\mathrm{p}}}(\mathbf{y}) = (\mathbf{y} - \mathbf{x}_{\mathrm{p}})/\left\|\mathbf{y} - \mathbf{x}_{\mathrm{p}}\right\|$ away from coincidence and zero at coincidence. For independent $\mathbf{x}_{\mathrm{r}} \sim Q$ and $\mathbf{x}_{\mathrm{s}} \sim P$, define $T = \beta_{\mathbf{x}_{\mathrm{p}}}(\mathbf{x}_{\mathrm{r}}) - \beta_{\mathbf{x}_{\mathrm{p}}}(\mathbf{x}_{\mathrm{s}})$ and $\mathbf{v}_P(\mathbf{x}_{\mathrm{p}}) = \mathbb{E}\left[T\right]$. Consider square-integrable source-sampling rules with auxiliary randomness independent of the sources, whose measurable form and auxiliary-randomness law do not depend on the unknown (P, Q); call such a rule universally unbiased when its expectation equals $\mathbf{v}_P(\mathbf{x}_{\mathrm{p}})$ for every (P, Q).

Proposition 5 (Source-law necessity and Rao–Blackwell optimality) . *No universally unbiased rule can omit either source law entirely. Among rules observing one independent draw from each law, T is the unique deterministic universally unbiased rule and minimizes covariance in the positive-semidefinite order, hence also mean-squared error, among randomized universally unbiased rules. For the average $\bar{T}_K$ of K independent copies,*

$$\mathbb{E}\left[\left\|\bar{T}_K - \mathbf{v}_P(\mathbf{x}_{\mathrm{p}})\right\|_2^2\right] = \frac{\operatorname{tr}\operatorname{Cov}_Q(\beta_{\mathbf{x}_{\mathrm{p}}}) + \operatorname{tr}\operatorname{Cov}_P(\beta_{\mathbf{x}_{\mathrm{p}}})}{K} \le \frac{2}{K}. \tag{43}$$

Proof. Without a draw from P (respectively Q), a rule's expectation cannot vary with that law, and two point masses with different bearings contradict universal unbiasedness. For any randomized one-draw-per-law rule A, taking $Q = \delta_{\mathbf{r}}$ and $P = \delta_{\mathbf{s}}$ forces $\mathbb{E}\left[A \mid \mathbf{x}_{\mathrm{r}}, \mathbf{x}_{\mathrm{s}}\right] = T$ pointwise. Hence $\operatorname{Cov}(A) = \operatorname{Cov}(T) + \mathbb{E}\left[\operatorname{Cov}(A \mid \mathbf{x}_{\mathrm{r}}, \mathbf{x}_{\mathrm{s}})\right]$, proving the optimality and deterministic uniqueness claims. Independence gives the variance identity, averaging supplies $1/K$, and each covariance trace is at most one because a bearing has norm at most one. □

Also $\left\|T\right\|_2 \le 2$ almost surely. With zero collision probability, the numerator in (43) equals $2 - \left\|\mathbb{E}_{\mathbf{x}_{\mathrm{r}}\sim Q}\left[\beta_{\mathbf{x}_{\mathrm{p}}}(\mathbf{x}_{\mathrm{r}})\right]\right\|_2^2 - \left\|\mathbb{E}_{\mathbf{x}_{\mathrm{s}}\sim P}\left[\beta_{\mathbf{x}_{\mathrm{p}}}(\mathbf{x}_{\mathrm{s}})\right]\right\|_2^2$. The source-law statement concerns universal dependence on both laws, not whether every randomized realization must query both.

F.5 Implementation and Interpretive Derivations

F.5.1 Denominator smoothing as a proper objective

Related work constructs differentiable approximations of negative-distance kernels for Wasserstein gradient flows (Rux et al., 2026); here we analyze the denominator form used in our implementation. For $\varepsilon > 0$, define $\psi_\varepsilon(r) = r - \varepsilon \log(1 + r/\varepsilon)$ and let $\mathcal{F}_\varepsilon$ be (14) with every distance r replaced by $\psi_\varepsilon(r)$.

Proposition 6 (Proper objective induced by denominator smoothing) . *Wherever the first variation is valid, the denominator-smoothed $\lambda = 1$ field satisfies*

$$\mathbf{v}_\varepsilon(\mathbf{x}) = \mathbb{E}_{\mathbf{x}_\mathrm{r}\sim Q}\left[\frac{\mathbf{x}_\mathrm{r} - \mathbf{x}}{\|\mathbf{x}_\mathrm{r} - \mathbf{x}\|_2 + \varepsilon}\right] - \mathbb{E}_{\mathbf{x}_\mathrm{s}\sim P}\left[\frac{\mathbf{x}_\mathrm{s} - \mathbf{x}}{\|\mathbf{x}_\mathrm{s} - \mathbf{x}\|_2 + \varepsilon}\right] = -\nabla_{\mathbf{x}} \frac{\delta \mathcal{F}_\varepsilon}{\delta P}(\mathbf{x}). \tag{44}$$

Moreover, $\mathcal{F}_\varepsilon(P) \geq 0$ and, under finite first moments, equality holds if and only if $P = Q$.

Proof. Since $\psi'_\varepsilon(r) = r/(r + \varepsilon)$, $\nabla_{\mathbf{x}} \psi_\varepsilon(\|\mathbf{x} - \mathbf{y}\|_2) = (\mathbf{x} - \mathbf{y})/(\|\mathbf{x} - \mathbf{y}\|_2 + \varepsilon)$, with zero gradient at coincidence. Applying this identity to the first variation proves (44); the frozen-target pullback gives the corresponding local parameter gradient.

For properness, write $\psi_\varepsilon(\sqrt{t}) = f_\varepsilon(t)$. Because $e^{-u\sqrt{t}}$ is completely monotone in t, the identity $(\sqrt{t} + \varepsilon)^{-1} = \int_0^\infty e^{-u\varepsilon} e^{-u\sqrt{t}}\, \mathrm{d}u$ shows that $f'_\varepsilon(t) = 1/[2(\sqrt{t} + \varepsilon)]$ is also completely monotone. Since $f_\varepsilon(0) = 0$, $f_\varepsilon(t)/t \to 0$ as $t \to \infty$, and f_ε is nonconstant, the Bernstein-function representation theorem (Schilling et al., 2012) gives a Lévy–Khintchine representation with neither constant nor linear term: $f_\varepsilon(t) = \int_{(0,\infty)} (1 - e^{-st})\, \nu_\varepsilon(\mathrm{d}s)$ for a nonzero positive Lévy measure ν_ε. Finite first moments and $0 \leq \psi_\varepsilon(r) \leq r$ justify applying Tonelli separately to the three nonnegative distance terms before combining them. Consequently, $2\mathcal{F}_\varepsilon(P) = \int_{(0,\infty)} \mathrm{MMD}^2_{k_s}(P, Q)\, \nu_\varepsilon(\mathrm{d}s)$ with Gaussian kernels $k_s(\mathbf{x}, \mathbf{y}) = e^{-s\|\mathbf{x}-\mathbf{y}\|_2^2}$. Each k_s is characteristic for $s > 0$, proving nonnegativity and identity of indiscernibles; finite first moments suffice because $0 \leq \psi_\varepsilon(r) \leq r$. □

The map $\mathbf{a} \mapsto \mathbf{a}/(\|\mathbf{a}\|_2 + \varepsilon)$ is globally $1/\varepsilon$-Lipschitz. Consequently, $\mathbf{v}_\varepsilon$ is $2/\varepsilon$-Lipschitz in its spatial argument and $1/\varepsilon$-Lipschitz in P under W_1; for fixed Q, the associated McKean–Vlasov characteristic dynamics are therefore well posed under finite first moments. The smoothed source estimator remains conditionally unbiased for $\mathbf{v}_\varepsilon$; hence the frozen-target, tracker-decomposition, and tracked-convergence arguments apply with $(\mathcal{F}, \mathbf{v})$ replaced by $(\mathcal{F}_\varepsilon, \mathbf{v}_\varepsilon)$ under the corresponding chain-rule and slope assumptions. Replacing the model–model coefficient by λ gives the analogous smoothed λ-weighted functional; before corridor conditioning, its negative-gradient field equals the corresponding source mean. The properness claim above is specific to $\lambda = 1$. Away from coincidence, the exact and smoothed source bearings differ in norm by $\varepsilon/(\|\mathbf{y} - \mathbf{x}\|_2 + \varepsilon)$; both are zero at coincidence. Therefore

$$\|\mathbf{v}_\varepsilon(\mathbf{x}) - \mathbf{v}(\mathbf{x})\|_2 \leq \mathbb{E}_{\mathbf{x}_\mathrm{r}\sim Q}\left[\frac{\varepsilon \mathbf{1}\{\mathbf{x}_\mathrm{r} \neq \mathbf{x}\}}{\|\mathbf{x}_\mathrm{r} - \mathbf{x}\|_2 + \varepsilon}\right] + \mathbb{E}_{\mathbf{x}_\mathrm{s}\sim P}\left[\frac{\varepsilon \mathbf{1}\{\mathbf{x}_\mathrm{s} \neq \mathbf{x}\}}{\|\mathbf{x}_\mathrm{s} - \mathbf{x}\|_2 + \varepsilon}\right] \longrightarrow 0 \quad (\varepsilon \downarrow 0). \tag{45}$$

The convergence follows by dominated convergence; if both source distances are at least $\delta > 0$, the bound is at most $2\varepsilon/(\delta + \varepsilon)$. Thus smoothing retains a proper objective while changing its geometry, and its field converges pointwise to the exact bearing field as $\varepsilon \downarrow 0$.

F.5.2 Near-endpoint surrogate control

Before tracker conditioning, the exact-bearing source mean of $\widehat{\mathbf{v}}_\lambda = \mathbf{b}_\mathrm{r} - \lambda \mathbf{b}_\mathrm{s}$ is the negative Wasserstein gradient field of

$$\mathcal{F}_\lambda(P; Q) = \mathbb{E}_{\mathbf{x}\sim P, \mathbf{y}\sim Q}\left[\|\mathbf{x} - \mathbf{y}\|\right] - \frac{\lambda}{2}\mathbb{E}_{\mathbf{x},\mathbf{x}'\sim P}\left[\|\mathbf{x} - \mathbf{x}'\|\right]. \tag{46}$$

At $\lambda = 1$, this equals $\frac{1}{2}D_E^2(P, Q)$ up to a Q-dependent constant. For $\lambda < 1$, $\mathcal{F}_\lambda = \mathcal{F}_1 + \frac{1-\lambda}{2}\mathbb{E}\left[\|\mathbf{x} - \mathbf{x}'\|\right]$; the positive pairwise term favors lower generated-sample dispersion and generally shifts the optimum away from Q. This concerns the projectile-conditioned field, whereas corridor conditioning targets a generally distinct conditional expectation.

Let ψ be either the exact potential $\psi(r) = r$ or the smoothed potential ψ_ε above, and define $\beta_\psi(\mathbf{a}) = \psi'(\|\mathbf{a}\|_2)\mathbf{a}/\|\mathbf{a}\|_2$ away from zero and $\beta_\psi(0) = 0$. In both cases $\|\beta_\psi\|_2 \le 1$, and

$$D_\psi^2(P,Q) = 2\mathbb{E}\left[\psi(\|\mathbf{x}-\mathbf{y}\|_2)\right] - \mathbb{E}\left[\psi(\|\mathbf{x}-\mathbf{x}'\|_2)\right] - \mathbb{E}\left[\psi(\|\mathbf{y}-\mathbf{y}'\|_2)\right],$$

$$\mathcal{F}_{\psi,\lambda}(P;Q) = \mathbb{E}\left[\psi(\|\mathbf{x}-\mathbf{y}\|_2)\right] - \frac{\lambda}{2}\mathbb{E}\left[\psi(\|\mathbf{x}-\mathbf{x}'\|_2)\right],$$

where $\mathbf{x},\mathbf{x}' \sim P$ and $\mathbf{y},\mathbf{y}' \sim Q$ are independent. The discrepancy D_ψ^2 is proper for either choice of ψ. Write $\mathbf{v}_{\psi,\lambda}(\mathbf{x}) = \mathbb{E}_{\mathbf{y}\sim Q}\left[\beta_\psi(\mathbf{y}-\mathbf{x})\right] - \lambda\mathbb{E}_{\mathbf{x}'\sim P}\left[\beta_\psi(\mathbf{x}'-\mathbf{x})\right]$ for the corresponding source-mean field before tracker conditioning.

The instant attraction-only endpoint. For $\lambda = 0$, define $f_{\psi,Q}(\mathbf{x}) = \mathbb{E}_{\mathbf{y}\sim Q}\left[\psi(\|\mathbf{x}-\mathbf{y}\|_2)\right]$ and $M_\psi(Q) = \arg\min_{\mathbf{x}\in\mathbb{R}^d} f_{\psi,Q}(\mathbf{x})$. If Q has a finite first moment, continuity and coercivity make $M_\psi(Q)$ nonempty for either admissible potential, and

$$\mathcal{F}_{\psi,0}(P;Q) = \int f_{\psi,Q}(\mathbf{x})\,P(\mathrm{d}\mathbf{x})$$

is minimized exactly by laws supported on $M_\psi(Q)$. For $\psi(r) = r$, this is the geometric-median set; if it is a singleton $\{\mathbf{x}^\star\}$, the unique minimizer is $\delta_{\mathbf{x}^\star}$ rather than Q, unless Q is already that point mass. Thus the instant $\rho = 0, \lambda = 0$ objective is generally not proper without generated-source repulsion. This characterization concerns the functional before corridor conditioning and does not characterize the tracked $\rho > 0$ update.

Proposition 7 (Controlled intra-source reweighting) . *Before corridor conditioning, the source-mean field $\mathbf{v}_{\psi,\lambda}$ obeys $\|\mathbf{v}_{\psi,\lambda}(\mathbf{x}) - \mathbf{v}_{\psi,1}(\mathbf{x})\|_2 \le 1-\lambda$. Moreover, if $\delta \ge 0$ and $\mathcal{F}_{\psi,\lambda}(P;Q) \le \mathcal{F}_{\psi,\lambda}(Q;Q) + \delta$, then*

$$D_\psi^2(P,Q) \le (1-\lambda)\mathbb{E}\left[\psi(\|\mathbf{y}-\mathbf{y}'\|_2)\right] + 2\delta. \tag{47}$$

Proof. The field difference is $(1-\lambda)\mathbb{E}_{\mathbf{x}'\sim P}\left[\beta_\psi(\mathbf{x}'-\mathbf{x})\right]$, proving the uniform bound. Also, $\mathcal{F}_{\psi,\lambda}(P;Q) = \frac{1}{2}D_\psi^2(P,Q) + \frac{1}{2}\mathbb{E}\left[\psi(\|\mathbf{y}-\mathbf{y}'\|_2)\right] + \frac{1-\lambda}{2}\mathbb{E}\left[\psi(\|\mathbf{x}-\mathbf{x}'\|_2)\right]$. Comparing this identity at P and Q and dropping the nonnegative model–model term proves (47). □

Consequently, if $\lambda_n \uparrow 1$, $\delta_n \to 0$, and $\mathcal{F}_{\psi,\lambda_n}(P_n;Q) \le \mathcal{F}_{\psi,\lambda_n}(Q;Q) + \delta_n$, then $D_\psi(P_n,Q) \to 0$; this includes global minimizers whenever they exist.

For the full mixed target, let $Y_{\psi,\lambda}$ denote the instant vector, retain the query $\mathbf{q}$ and history $\mathcal{H}$ from Sec. F.2.2, and write $m_\lambda(\mathbf{q}) = \mathbb{E}\left[Y_{\psi,\lambda} \mid \mathbf{q},\mathcal{H}\right]$ and $\bar{\mathbf{v}}_{\psi,\lambda}(\mathbf{x}) = \mathbb{E}\left[Y_{\psi,\lambda} \mid \mathbf{x},\mathcal{H}\right] = \mathbf{v}_{\psi,\lambda}(\mathbf{x})$. Define the tracker excess risk $E_{\boldsymbol{\phi}} = \mathbb{E}\left[\|h_{\boldsymbol{\phi}}(\mathbf{q}) - m_\lambda(\mathbf{q})\|_2^2 \mid \mathcal{H}\right]$ and the oracle corridor mismatch $C_\lambda = \|\mathbb{E}\left[m_\lambda(\mathbf{q}) \mid \mathbf{x},\mathcal{H}\right] - \bar{\mathbf{v}}_{\psi,\lambda}(\mathbf{x})\|_{L^2(P)}$. In this section, set $G_\rho = (1-\rho)Y_{\psi,\lambda} + \rho h_{\boldsymbol{\phi}}(\mathbf{q})$, and let $u_\rho(\mathbf{x}) = \mathbb{E}\left[G_\rho \mid \mathbf{x},\mathcal{H}\right]$ be the mean mixed field.

Proposition 8 (Stability of the deployed mixed field) . *The deviation from the proper endpoint field satisfies*

$$\|u_\rho - \mathbf{v}_{\psi,1}\|_{L^2(P)} \le \underbrace{(1-\lambda) + \rho\left(C_\lambda + \sqrt{E_{\boldsymbol{\phi}}}\right)}_{\Delta}. \tag{48}$$

Proof. The exact decomposition is $u_\rho - \mathbf{v}_{\psi,1} = (\bar{\mathbf{v}}_{\psi,\lambda} - \mathbf{v}_{\psi,1}) + \rho\mathbb{E}\left[h_{\boldsymbol{\phi}}(\mathbf{q}) - m_\lambda(\mathbf{q}) \mid \mathbf{x},\mathcal{H}\right] + \rho(\mathbb{E}\left[m_\lambda(\mathbf{q}) \mid \mathbf{x},\mathcal{H}\right] - \bar{\mathbf{v}}_{\psi,\lambda})$. Apply conditional L^2 contraction and Prop. 7. □

At $\lambda = 1$, the query is the projectile and $C_1 = 0$; for $\lambda < 1$, C_λ isolates the bias caused by conditioning on the corridor rather than the projectile. Let P_t solve the continuity equation with velocity $u_{\rho,t}$, let Δ_t denote the right-hand side of (48), and write $\mathcal{F}_\psi(P) = \frac{1}{2}D_\psi^2(P,Q)$. Under the corresponding chain rule and linear slope $\|\mathbf{v}_{\psi,1,t}\|_{L^2(P_t)}^2 \ge c\mathcal{F}_\psi(P_t)$, Young's inequality yields

$$\mathcal{F}_\psi(P_t) \le e^{-ct/2}\mathcal{F}_\psi(P_0) + \frac{1}{2}\int_0^t e^{-c(t-s)/2}\Delta_s^2\,\mathrm{d}s. \tag{49}$$

Hence bounded $\Delta_t \to 0$ gives $D_\psi(P_t, Q) \to 0$, while $\sup_t \Delta_t \le \bar{\Delta}$ gives $\limsup_{t\to\infty} \mathcal{F}_\psi(P_t) \le \bar{\Delta}^2/c$. If instead $\Delta_t \le \gamma \|\mathbf{v}_{\psi,1,t}\|_{L^2(P_t)}$ for some $\gamma < 1$, the argument of Thm. 3 gives exact convergence under its remaining assumptions. Thus (47) controls the weighted objective, while (49) quantifies stability of the deployed flow around the proper endpoint.

F.5.3 Line-integral interpretation of the displacement tracker

For the modified displacement analogue in Sec. E.1, let $\boldsymbol{v}_{\boldsymbol{\phi}} = \nabla u_{\boldsymbol{\phi}}$, $\Delta = \mathbf{x}_\mathrm{r} - \mathbf{x}_\mathrm{p}$, and $\mathbf{x}_\alpha = (1-\alpha)\mathbf{x}_\mathrm{p} + \alpha \mathbf{x}_\mathrm{r}$. Use the second-moment, differentiability, and integrability assumptions stated with (13); Cauchy–Schwarz then makes the cross term integrable and justifies exchanging its expectation and the α integral. The fundamental theorem for line integrals gives

$$\int_0^1 \nabla_{\mathbf{x}} u_{\boldsymbol{\phi}}(\mathbf{x}_\alpha, \mathbf{c})^\top \Delta \, \mathrm{d}\alpha = u_{\boldsymbol{\phi}}(\mathbf{x}_\mathrm{r}, \mathbf{c}) - u_{\boldsymbol{\phi}}(\mathbf{x}_\mathrm{p}, \mathbf{c}). \tag{50}$$

Expanding $\mathbb{E}\left[\|\nabla u_{\boldsymbol{\phi}}(\mathbf{x}_\alpha, \mathbf{c}) - \Delta\|_2^2\right]$ for $\alpha \sim \mathcal{U}(0,1)$ and applying this identity gives (13); the remaining $\mathbb{E}\left[\|\Delta\|_2^2\right]$ is constant for fixed generator parameters. Hence displacement regression maximizes the real–generated potential gap minus one half of the corridor-averaged squared gradient norm. The infimum over conservative fields is characterized by the L^2 projection of $\mathbb{E}[\Delta \mid \mathbf{x}_\alpha, \mathbf{c}]$ onto their closure under the corridor measure; a finite network is further restricted by its parameterization. This is not an identity for the deployed unit-bearing update, and a general vector tracker retains the conditional-expectation decomposition but not the scalar line-integral interpretation.

G Limitations

First, population-flow convergence assumes the exact $\lambda = 1$ field, chain-rule regularity, a slope inequality, and relative tracking; finite-generator convergence instead assumes a smooth exact objective, realizability, gradient dominance, relative bias, bounded variance, and diminishing steps. Neither establishes convergence for general neural generators trained by SGD. Second, vector estimates may be noisy near coincident particles; at $\lambda = 1$, fixed denominator smoothing stabilizes them and retains a proper smoothed objective, but changes its geometry. Third, the efficiency and quality of TBSM from random initialization at ImageNet scale remain untested. Fourth, even at $\lambda = 1$, representation-space properness concerns only projected distributions, so full image matching depends on the information retained by the encoder. Fifth, the benchmark establishes sample quality at NFE $= 1$, but not compute-matched training efficiency against mature diffusion and autoregressive systems. Finally, in strongly conditional settings where each condition is paired with only one real sample, the empirical conditional energy objective may encourage matching that single paired sample unless diversity is supplied through data augmentation, shared statistical structure across conditions, or additional regularization. The Qwen-Image-20B benchmark scores do not establish diversity across multiple samples for the same prompt; a dedicated conditional-diversity evaluation remains necessary.