GRADIENT FLOW DRIFTING: ONE-STEP GENERATIVE MODELING WITH KDE DIVERGENCE FLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce *Gradient Flow Drifting* (GFD), a unified, directly estimable energy-defined framework for one-step implicit generative modeling. GFD treats state-space geometry, kernel representation, and divergence energy as explicit modeling choices and maps them to particle dynamics and generator training through a common variational construction. The central principle is to take the Wasserstein gradient of the KDE-defined energy with respect to the clean model distribution and amortize the resulting particle updates into a neural generator. For admissible kernel-density f-divergences, we derive the first variation, clean-particle fields, energy dissipation along classical flows, and distributional identification under injective smoothing. The framework also incorporates a quadratic distance between kernel densities as a density-ratio-free energy and supports nonnegative energy mixtures through the corresponding linear combination of particle fields. Gaussian and von Mises–Fisher kernels give explicit Euclidean and spherical instances. Sample-based field estimation and stop-gradient regression provide a common training procedure while preserving single-pass inference. Experiments on encoder-free synthetic distributions, spherical directional data, and high-dimensional DNA sequences, alongside encoder-based CIFAR-10 and ImageNet-256 generation, demonstrate the effectiveness and versatility of representative GFD instances and analyze the behavior of different divergence choices.

1 INTRODUCTION

One-step generators need a distribution-level training signal when both the target p and model q_θ are accessible through samples. GANs learn a discriminator, while diffusion and flow models learn time-dependent denoising or transport objectives (Goodfellow et al., 2014; Ho et al., 2020; Lipman et al., 2023). Drifting learns to reproduce particle movements within a neural generator (Deng et al., 2026). This raises a constructive question: *can a family of one-step models be specified through its state-space geometry, kernel representation, and distributional energy?*

Gradient Flow Drifting (GFD) organizes a directly estimable family of one-step models in which an explicit distributional energy determines the clean-particle learning signal:

$$\underbrace{\mathcal{G}(q;p)}_{\text{energy specification}} \longmapsto \underbrace{\mathbf{u}_{\mathcal{G}}[q]}_{\text{clean-measure flow}} \longmapsto \underbrace{\widehat{\mathbf{u}}_{\mathcal{G}}}_{\text{sample estimator}} \longmapsto \underbrace{\mathrm{sg}[x^+]}_{\text{neural target}} . \tag{1}$$

Building on the energy-to-flow principle, GFD's distinctive contribution is a direct KDE route: general divergence responses, the adjoint clean-particle field, its sample estimator, and the stopped neural target arise within one explicit construction. The energy determines the response to distributional error; the kernel and geometry determine particle interactions. Differentiating the energy with respect to the *clean model measure* gives a Wasserstein descent field. Sample-based estimates define stopped particle targets that are amortized into G_θ, with $q_\theta = (G_\theta)_{\#}p_\epsilon$. Training repeatedly updates the generated distribution, while inference remains one forward pass. The same construction determines both the population dynamics and the particle targets used by the training algorithm.

The framework provides concrete modeling capabilities. Direct KDE f-divergences change the density-ratio response through f, including KL, reverse KL, Pearson, Hellinger, and mixtures.

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.

Gaussian and von Mises–Fisher (vMF) kernels give Euclidean and spherical interactions. Nonnegative energy combinations yield linear combinations of their fields, and a quadratic distance between KDEs supplies the density-ratio-free KDE-L^2 component. These choices share one estimation and generator-training interface.

Our contributions center on the following two advances.

A directly estimable energy-defined family for one-step training. We develop the direct KDE family $q \mapsto D_f(\mathcal{S}_k q \| \mathcal{S}_k p)$, in which geometry, a normalized smoothing kernel, and a divergence response yield an explicit clean-particle field and sample estimator. The construction covers general KDE f-divergences, the quadratic KDE energy, and nonnegative compositions, with exact first variations, classical-flow dissipation, and distributional identification under injective smoothing. Kernel-query estimation and stopped-target regression provide a common interface from these population fields to one-step generators.

Controlled empirical gains and effective energy composition. GFD+mix achieves the lowest FID at every reported checkpoint in the controlled CIFAR-10 comparison against Drifting and W-Flow. On ImageNet-256, GFD-L leads the compared directly trained models in each of the five family-balanced discrepancies, and retains the best aggregate after removing any one evaluation family. A five-seed Swiss-roll study shows that KDE-L^2 + forward KL ties for the lowest mean discrepancy and has the smallest observed standard deviation. Encoder-free spherical, Earth-event, and high-dimensional DNA experiments further demonstrate the framework's scope, with one-step DNA generation matching the highest reported 6-mer correlation in the comparison.

2 Related Work

Distributional objectives and one-step training. MMD-GAN and Sinkhorn objectives train generators with sample discrepancies (Li et al., 2017; Genevay et al., 2018); consistency models and Generator Matching provide other routes to direct generation (Song et al., 2023; Holderrieth et al., 2025). GFD shares distribution matching with these approaches and stopped-target amortization with Drifting (Deng et al., 2026).

Clean-measure gradient flows and drifting. Gretton et al. (2026) distinguish pointwise score fields of smoothed densities from clean-measure variations of smoothed energies, including KDE-KL. W-Flow instantiates one-step training with a Sinkhorn-divergence Wasserstein flow (Han et al., 2026a). KGD develops kernel-score updates on Euclidean and manifold domains; its smoothed-law descent and clean-particle application are distinct from the clean-measure KDE-divergence flow studied here (Esteban-Casadevall et al., 2026). GFD focuses on a direct, sample-computable KDE f-family: scalar density-ratio responses pass through the adjoint smoothing operator to clean-particle fields and kernel-query estimators, with a compatible quadratic KDE component and energy composition. Appendix M compares the constructions and their field equations.

3 Gradient Flow Drifting

3.1 Unified model specification

Definition 3.1 (GFD instance)**.** *A GFD model specifies a state space (M, g), a target p, a kernel representation $\mathcal{K}$, a differentiable energy $\mathcal{G}(q; p)$, and a generator G_θ with law q_θ. Its population field is $\mathbf{u}_{\mathcal{G}}[q] = -\nabla_g(\delta\mathcal{G}/\delta q)$. An implementation chooses a sample estimator, a particle step, and stopped-target regression.*

The definition specifies a model through three linked choices. The optimization space M determines admissible particle positions and the metric used to move them. The kernel representation $\mathcal{K}$ maps probability measures to smooth densities, and the energy $\mathcal{G}$ measures their discrepancy. The variational field then determines the particle target used by the training algorithm; changing any of these choices produces another instance of the same framework.

The optimization space is unrestricted by the definition: it may be the data domain, a manifold, or the feature space of a frozen encoder. For a data-space generator g_θ and encoder E, feature-space

Table 1: Divergence responses with $r = s_q/s_p$. All use the same adjoint construction.

Energy	$f(r)$	$f'(r)$
$\mathrm{KL}(s_q\|s_p)$	$r\log r$	$\log r + 1$
$\mathrm{KL}(s_p\|s_q)$	$-\log r$	$-r^{-1}$
Pearson $\chi^2(s_q\|s_p)$	$(r-1)^2$	$2(r-1)$
Squared Hellinger	$(\sqrt{r}-1)^2$	$1-r^{-1/2}$

optimization sets $G_\theta = E \circ g_\theta$ and $p = E_\# P_{\mathrm{data}}$. If E is injective on the data domain with a measurable inverse on its image, equality of feature distributions implies equality of the original distributions. Together with injective kernel smoothing, this transfers distributional identification back to data space (Appendix B.4). For a general non-injective feature map, matching and identification concern the induced feature distribution; injectivity of the frozen image encoders used experimentally is not assumed.

3.2 Variational construction and energy families

Let (M, g) be a boundaryless Riemannian manifold and $\varphi(x, y) > 0$ a kernel normalized in y. Define its density and adjoint averaging operators by

$$s_\mu(y) = (\mathcal{S}_k\mu)(y) = \int_M \varphi(x,y)\,\mathrm{d}\mu(x), \tag{2}$$

$$(\mathcal{A}_k a)(x) = \int_M a(y)\varphi(x,y)\,\mathrm{dvol}_g(y). \tag{3}$$

For convex differentiable f with $f(1) = 0$, the energy is (Csiszár, 1967)

$$\mathcal{G}_f(q;p) = D_f(\mathcal{S}_k q\|\mathcal{S}_k p) = \int_M s_p(y) f(s_q(y)/s_p(y))\ \mathrm{dvol}_g(y). \tag{4}$$

Assumption 3.2 (Admissibility)**.** *The densities are positive, all displayed integrals are finite, differentiation can be interchanged with integration, and integration by parts has no boundary contribution. A closed manifold, positive C^2 kernel, and $f \in C^2$ suffice for the KDE formulas. For the quadratic KDE energy, kernel sections are square-integrable and their induced kernel has integrable first derivatives. Appendix A gives the noncompact conditions.*

Theorem 3.3 (Clean-measure KDE f-flow)**.** *Under Assumption 3.2, up to an additive constant,*

$$\frac{\delta \mathcal{G}_f}{\delta q}(x) = \mathcal{A}_k[f'(s_q/s_p)](x), \tag{5}$$

$$\begin{aligned}\mathbf{u}_f[q](x) &= -\nabla_x \mathcal{A}_k[f'(s_q/s_p)](x)\\ &= -\int_M f'(s_q(y)/s_p(y))\,\nabla_x\varphi(x,y)\,\mathrm{dvol}_g(y).\end{aligned} \tag{6}$$

Along a classical solution of $\partial_t q_t + \mathrm{div}(q_t \mathbf{u}_f[q_t]) = 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{G}_f(q_t;p) = -\int_M \|\mathbf{u}_f[q_t](x)\|_g^2\,\mathrm{d}q_t(x) \le 0. \tag{7}$$

If f is strictly convex and $\mathcal{S}_k$ is injective on probability measures, $\mathcal{G}_f(q;p) = 0$ exactly when $q = p$. On connected M, the global field identity $\mathbf{u}_f[q] \equiv 0$ also holds exactly when $q = p$.

The proof is in Appendix B. The adjoint $\mathcal{A}_k$ is the chain rule through kernel smoothing: it maps the query-space discrepancy response back to clean-particle locations. Thus the energy, its field, and distributional identification belong to one variational construction.

Table 1 makes the density-ratio response an explicit modeling choice. Changing f changes how discrepancies are weighted before adjoint averaging; adding $a(r-1)$ does not change the divergence or its field. This supplies distinct responses without fitting a separate critic.

KDE-L^2: a quadratic KDE energy. The same construction yields a density-ratio-free model by choosing

$$\mathcal{G}_2(q;p) = \int_M (s_q(y) - s_p(y))^2 \,\mathrm{dvol}_g(y), \qquad k_\varphi(x,x') = \int_M \varphi(x,y)\varphi(x',y)\,\mathrm{dvol}_g(y). \tag{8}$$

Expanding the squared density difference and applying Fubini gives $\mathcal{G}_2 = \iint k_\varphi(x,x')\,\mathrm{d}(q-p)(x)\,\mathrm{d}(q-p)(x')$. Its first variation is $2\mathcal{A}_k[s_q - s_p]$, so the KDE-$L^2$ field is

$$\mathbf{u}_2[q](x) = -2\nabla_x \mathcal{A}_k[s_q - s_p](x) = 2\int_M \nabla_1 k_\varphi(x,x')\,\mathrm{d}(p-q)(x'). \tag{9}$$

This model compares KDEs directly and admits a pairwise estimator without density ratios. Injective smoothing gives $\mathcal{G}_2 = 0$ exactly when $q = p$. Appendix C proves the identities and dissipation, and relates the induced kernel k_φ to the classical RKHS discrepancy (Arbel et al., 2019).

3.3 Geometry and energy composition

Kernel choice specifies both the smoothing operator and the query-noise distribution: the conditional query law has density $\varphi(x,\cdot)$ with respect to the volume measure of M. Appendix E.1 gives a selection procedure, normalization rules, and exact samplers. Two instances illustrate the construction.

Gaussian / Euclidean instance. For $M = \mathbb{R}^d$ and $\varphi_h(x,y) = \mathcal{N}(y; x, h^2 I)$, the kernel derivative gives

Corollary 3.4.

$$\mathbf{u}_f[q](x) = -\mathbb{E}_{Y\sim\mathcal{N}(x,h^2 I)}\left[f'(s_q(Y)/s_p(Y))\,(Y-x)/h^2\right]. \tag{10}$$

For KL, integration by parts equivalently gives

$$\mathbf{u}_{\mathrm{KL}}[q](x) = \mathbb{E}_{Z\sim\mathcal{N}(0,h^2 I)}[\nabla \log s_p(x+Z) - \nabla \log s_q(x+Z)]. \tag{11}$$

vMF / spherical instance. For $M = \mathbb{S}^{d-1}$ and $\varphi_\kappa(x,y) = C_d(\kappa)e^{\kappa x^\top y}$, the intrinsic kernel score is $\kappa(y - (x^\top y)x)$, giving

Corollary 3.5.

$$\mathbf{u}_f[q](x) = -\kappa\,\mathbb{E}_{Y\sim\mathrm{vMF}(x,\kappa)}\left[f'(s_q(Y)/s_p(Y))\left(Y - (x^\top Y)x\right)\right]. \tag{12}$$

This field is tangent to the sphere; a retraction implements its particle step. Both instances change interaction geometry while retaining the same divergence response and variational principle. Appendix E gives the derivations and further kernels.

Proposition 3.6 (Energy composition)**.** *For admissible energies on a common state space and fixed $\alpha_j \geq 0$,*

$$\mathcal{G}_{\mathrm{mix}} = \sum_j \alpha_j \mathcal{G}_j, \quad \mathbf{u}_{\mathrm{mix}} = \sum_j \alpha_j \mathbf{u}_j, \quad \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{G}_{\mathrm{mix}}(q_t) = -\int \|\mathbf{u}_{\mathrm{mix}}\|_g^2 \,\mathrm{d}q_t \tag{13}$$

along a classical mixed flow.

Proposition 3.6 concerns fixed nonnegative weights and population component fields before numerical scale balancing. As in drifting-style implementations, estimated fields may be balanced before finite target construction; Appendix K.1 specifies this step. KDE f-responses and KDE-L^2 can therefore be composed before generator training. With a shared KDE kernel and target, a mixture of f-divergences has response $f_{\mathrm{mix}} = \sum_j \alpha_j f_j$. Adding $\lambda\mathcal{G}_2$ further supplies a density-ratio-free contribution through $\lambda\mathbf{u}_2$ (Appendix G).

3.4 From particle fields to one-step generators

Empirical densities and the query estimator. Given independent real references $X_j^+ \sim p$ and model references $X_j^- \sim q$, define

$$\widehat{s}_p(y) = \frac{1}{n_p}\sum_{j=1}^{n_p}\varphi(X_j^+, y), \quad \widehat{s}_q(y) = \frac{1}{n_q}\sum_{j=1}^{n_q}\varphi(X_j^-, y), \quad \widehat{r}(y) = \frac{\widehat{s}_q(y)}{\widehat{s}_p(y)}. \tag{14}$$

For a clean particle x, sample $Y_\ell \sim \varphi(x, \cdot)$ independently of these references and compute

$$\widehat{\mathbf{u}}_f(x) = -\frac{1}{L}\sum_{\ell=1}^{L} f'(\widehat{r}(Y_\ell))\, b_\varphi(x, Y_\ell), \qquad b_\varphi(x, y) = \nabla_x \log \varphi(x, y). \tag{15}$$

This is a direct computation from kernel values, their derivatives, and scalar responses. Gaussian queries use $Y = x + h\epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, with $b_\varphi = \epsilon/h$; vMF queries use the intrinsic score in Eq. (12). In both cases, shared normalizing constants cancel in $\widehat{r}$. The KDE-L^2 estimator instead averages pairwise derivatives of k_φ (Appendix C).

Reference-sample bias. The empirical densities are unbiased at a fixed query, but their ratio is generally biased. For fixed kernel parameters, $n_p, n_q \asymp n$, independent reference batches, denominators bounded away from zero, and the moment conditions in Appendix I.1,

$$\mathbb{E}[\widehat{r}(y)] - r(y) = \frac{r(y)\,\mathrm{Var}_{X\sim p}[\varphi(X, y)]}{n_p\, s_p(y)^2} + o(n^{-1}). \tag{16}$$

Thus the leading ratio bias is $O(n^{-1})$. A second-order expansion of $f'(\widehat{r})$ gives the same order for the response and, under integrable domination, the field. Appendix I.1 derives the coefficient, the corresponding result for correlated references, and the separate roles of reference count and query count. Leave-one-out references remove a particle's self-term when the current model batch is reused.

Particle targets and generator updates. For $x = G_\theta(\epsilon)$, form $x^+ = \mathrm{Retr}_x(\eta\widehat{\mathbf{u}}_{\mathcal{G}}(x))$ and train by

$$\mathcal{L}_{\mathrm{GFD}}(\theta) = \mathbb{E}_\epsilon\left[d_M^2\big(G_\theta(\epsilon), \mathrm{sg}[x^+]\big)\right]. \tag{17}$$

In Euclidean optimization coordinates, the stopped gradient is $-2\eta\mathbb{E}[J_\theta G_\theta(\epsilon)^\top \widehat{\mathbf{u}}_{\mathcal{G}}]$. For feature-space optimization, $G_\theta = E \circ g_\theta$ transfers the same field through the encoder and generator Jacobians. KDE f-energies, KDE-L^2, and mixed fields therefore share one training algorithm (Appendix H); inference evaluates the data-space generator once. The population identities define the underlying energy-derived fields; finite-reference estimation, field-scale balancing, finite particle targets, and stopped neural regression form the implementation layers.

Conditional generation and guidance. Conditioning replaces p, q by p_c, q_c. For KL, the query-density construction gives $\mathbf{u}_{\mathrm{cfg}} = (1 + w)\mathbf{u}_c - w\mathbf{u}_\varnothing$, with both component fields using the same model KDE. As in ODE/SDE-based generators, including flow matching and diffusion, classifier-free guidance combines conditional and unconditional fields; this operation alone does not identify a prescribed tilted endpoint distribution (Ho & Salimans, 2022; Lipman et al., 2023; Bradley & Nakkiran, 2025). For GFD, the corresponding question is whether the guided query density is the smoothing of a clean target. Appendix F derives the guidance rule and explains this shared distinction between field interpolation and target-law identification.

4 Experiments

The experiments test three capabilities: meaningful divergence choices, operation across geometries and data types, and scalable one-step generation. The cross-method comparisons use GFD-KL and GFD+mix, whose underlying population energy equally combines reverse KL and half-scaled Pearson χ^2; Appendix K.1 specifies the finite-step component-scale normalization. A dedicated five-seed study further compares six individual and mixed energies. Toy, sphere, Earth-event, and DNA fields act directly on native coordinates without a learned feature encoder. CIFAR-10 and ImageNet instead use feature-space objectives. All GFD results use the corrected clean-measure field. Detailed protocols and reference tables are in Appendices K and J.

4.1 Divergence responses under controlled comparisons

Native-coordinate toy tests. The Euclidean comparison uses the original Drifting demo's samplers and a shared MLP, target batches, and 5,000-update budget (Deng et al., 2026). The sphere uses the released KGD tasks, backbone, and intrinsic operations (Esteban-Casadevall et al., 2026). Table 2 reports the minimum raw-model SW_2 recorded during training under the common within-task

Table 2: Toy SW_2: best recorded over 5,000 updates. Dashes denote unreported comparisons.

	Euclidean		Sphere	
Method	Checkerboard	Swiss roll	Checkerboard	Swiss roll
KGD	0.0348	0.0116	**0.0126**	0.0195
Drifting	**0.0148**	0.0112	–	–
W-Flow	0.0247	0.0355	–	–
GFD-KL	0.0207	0.0159	0.0168	0.0153
GFD+mix	0.0199	**0.0097**	0.0163	**0.0126**

Table 3: Divergence and energy composition on Euclidean Swiss roll over five training seeds. SW_2 is lower-is-better; best and worst denote the lowest and highest seed-level values. Bold marks the smallest mean, standard deviation, and extrema. Further evaluation details are in Appendix K.2.

Energy	SW_2 (mean $\pm$ std.) $\downarrow$	Best $\downarrow$	Worst $\downarrow$
Forward KL	0.0100 ± 0.0025	0.0074	0.0130
Reverse KL	0.0107 ± 0.0015	0.0085	0.0123
Pearson χ^2	$\mathbf{0.0099} \pm 0.0043$	**0.0063**	0.0170
KDE-L^2	0.0107 ± 0.0016	0.0080	0.0123
Reverse KL + χ^2	0.0102 ± 0.0022	0.0074	0.0132
KDE-L^2 + forward KL	$\mathbf{0.0099 \pm 0.0008}$	0.0088	**0.0107**

selection rule. GFD+mix improves on GFD-KL on both Euclidean and spherical tasks. Drifting and KGD lead on their respective checkerboards; GFD+mix is best on both Swiss rolls. These results show how changing the response improves distribution matching within a fixed geometric construction.

Divergence and energy composition across seeds. We compare six energy choices on Euclidean Swiss roll using five training seeds per choice (Table 3). KDE-L^2 + forward KL ties Pearson χ^2 for the lowest mean SW_2 (0.0099), while giving the smallest standard deviation (0.0008) and the lowest worst-seed SW_2 (0.0107). The mixed objective improves on both forward KL and KDE-L^2 alone in mean discrepancy and observed variability. This provides direct evidence that composing the density-ratio-free and KL fields can combine accurate distribution matching with consistent performance across seeds. The complete evaluation details are in Appendix K.2.

Controlled CIFAR-10: training-method effectiveness. We test whether GFD produces better generation quality than comparable algorithms under the same experimental conditions. GFD, Drifting, and W-Flow share the feature extractor, image preprocessing, DriftDiT generator, optimizer, batch size, training schedule, and evaluation code, while using their respective released training algorithms. GFD+mix achieves lower FID at every reported checkpoint, and GFD-KL achieves the highest IS after 10k steps (Figure 1). The consistent FID advantage across the reported checkpoints demonstrates the effectiveness of the complete GFD training design under the shared representation, backbone, preprocessing, training schedule, and evaluation protocol.

4.2 Geometry and high-dimensional encoder-free generation

Earth events. On Volcano, Earthquake, Fire, and Flood, fields act directly on spherical coordinates. Methods follow the KGD benchmark split, backbone, budget, and evaluation code (Esteban-Casadevall et al., 2026; Mathieu & Nickel, 2020). GFD-KL gives the lowest Volcano discrepancy (0.0751), while GFD+mix gives the lowest Flood discrepancy (0.050); KGD leads on Earthquake and Fire (Table 4). These different preferences illustrate the role of the divergence response on a fixed geometry.

Promoter DNA. GFD trains on square-root categorical coordinates for length-1024 sequences over A, C, G, T (Appendix K.3). There is no pretrained sequence encoder, learned feature extractor, or compressed latent representation between the sequence and the GFD objective. We use vMF and

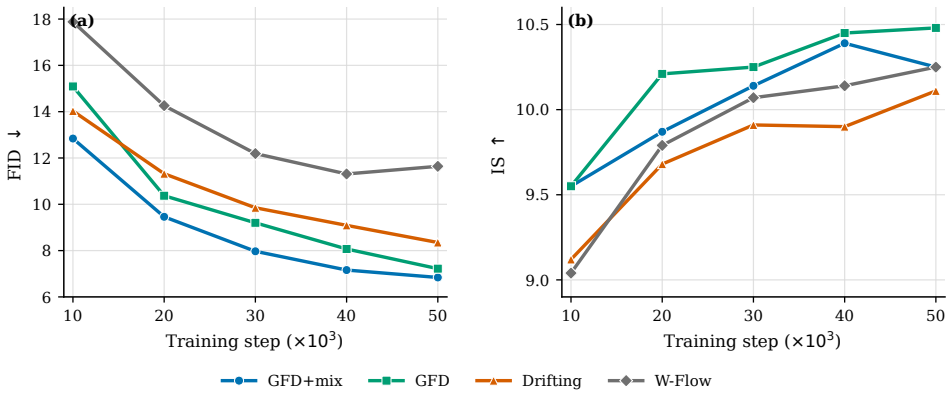


Figure 1: Controlled CIFAR-10 with shared features, preprocessing, backbone, and training schedule; methods retain their own released training algorithms. GFD+mix has the lowest reported FID throughout the comparison.

product-vMF kernels to match single-component and grouped joint distributions at sequence scales $[1, 4, 16, 64, 256, 1024]$, including the complete sequence scale. GFD+mix attains a 6-mer correlation of 0.96 at one NFE, matching Eulerian RMF and exceeding the reported Drifting and KGD values (Table 5). Thus encoder-free operation extends beyond low-dimensional examples.

Table 4: Earth-event kernel discrepancy.

Method	Volcano ↓	Earthquake ↓	Fire ↓	Flood ↓
Drifting	0.146	0.044	0.036	0.064
KGD	0.112	**0.037**	**0.029**	0.053
W-Flow	0.095	0.039	0.032	0.069
GFD-KL (ours)	**0.0751**	0.046	0.034	0.076
GFD+mix (ours)	0.121	0.051	0.035	**0.050**

Table 5: Promoter DNA sequence generation metrics.

Method	NFE ↓	6-mer Corr. ↑
Fisher FM (Davis et al., 2024)	100	0.96
Drifting Model (Deng et al., 2026)	1	0.88
KGD (Esteban-Casadevall et al., 2026)	1	0.91
Eulerian RMF (Woo et al., 2026)	1	0.96
GFD+mix (ours)	1	**0.96**

4.3 Scaling One-Step Image Generation

Multi-level feature-space construction. GFD uses frozen DINOv3 ViT-B/16 features (Siméoni et al., 2025) from layers $3, 6, 9, 12$, with CLS tokens and patch groups at each depth. ImageNet additionally uses ConvNeXt-B feature maps at resolutions 32, 16, and 8. Normalized features define vMF and product-vMF components across spatial scales and kernel temperatures; each component supplies a stopped target and loss (Appendix K.3).

Conventional image metrics and training budgets. Scaled GFD+mix achieves CIFAR-10 FID 2.04 at one NFE, compared with reported 2.51 for iCT-deep and 2.92 for MeanFlow among direct non-adversarial one-step references (Song & Dhariwal, 2024; Geng et al., 2025); Table 10 retains the full comparison across training regimes. On ImageNet, scaling GFD-B to GFD-L improves FID from 3.14 to 2.34 (Table 6). GFD-L does not obtain the lowest conventional Inception FID among the directly

Table 6: ImageNet-256 at one NFE. Baseline FID values are from the original papers; GFD values are from the reported GFD runs. IS is independently recomputed. The ratio panel in Table 7 is a separate checkpoint evaluation, not a rescaling of this FID column. Full configurations and citations are in Table 11.

Method	Training regime	Epochs (steps)	FID ↓	IS ↑
W-Flow-B	direct	1280 (200k)	1.52	284.44
W-Flow-L	direct	1280 (200k)	1.35	294.56
Drifting-L	direct	640 (100k)	1.61	346.10
pMF-L/16	direct	320 (400k)	2.52	299.43
GFD-B	direct	384 (120k)	3.14	293.18
GFD-L	direct	384 (120k)	2.34	315.69
AMFD-U (pMF-L)	post-training	+100 (125k)	0.86	343.58

trained references; the multi-representation analysis below provides a complementary distribution-matching evaluation. These results demonstrate generation quality at scale, complementing the shared-encoder CIFAR evidence for training-method effectiveness.

GFD combines single-pass inference with a compact training representation: its two frozen encoders contain 173.2M parameters, compared with the 2.69B representation model used by Drifting and W-Flow. GFD-B/L use 384 training epochs, while Drifting-L uses 640 and W-Flow-B/L use 1280. Together with FID 2.34 and the multi-representation results below, these configurations demonstrate scalable one-step generation with strong distribution matching and a smaller frozen representation model.

4.4 Distribution matching across evaluation representations

FID evaluates Gaussian-moment agreement in supervised Inception-v3 features. In feature-space generative modeling, this introduces a critical evaluation confound: optimizing a generator against representations fine-tuned for ImageNet classification directly injects class-discriminative label supervision. As established by Kynkäänniemi et al. (2023) and empirically corroborated by Deng et al. (2026, Table 3), matching such class-discriminative cues substantially decrease Inception FID without a commensurate gain in perceptual fidelity, and can even invert the relationship between FID and visual quality (Yang et al., 2026). Therefore, GFD operates strictly on label-free, self-supervised representations (DINOv3) to eliminate this supervisory bias. Consequently, conventional Inception FID conflates generative fidelity with class-label alignment. Complementing our controlled comparison on CIFAR-10 with related methods, evaluating GFD on ImageNet-256 under a strictly uncompromised, label-free protocol makes assessment across diverse, held-out representations essential for a fair and holistic distribution-matching benchmark.

We evaluate both representation diversity and discrepancy diversity to test whether distribution-matching gains extend beyond a classifier-aligned proxy. Following the motivation of CMMD and DINO-FD (Jayasumana et al., 2024; Stein et al., 2023), we independently evaluate released baseline and GFD checkpoints in six encoder families not used to train GFD: SigLIP2, ConvNeXtV2-FCMAE, Inception-v3, DreamSim, EUPE, and Pixio. Combining these representations with multiple discrepancy measures exposes differences hidden by a single feature space or its Gaussian-moment summary, while retaining Inception as one member of the evaluation panel.

For generated, training, and validation distributions P_G, P_T, P_V, define

$$r_{\phi,D}(G) = \frac{D_\phi(P_G, P_T)}{D_\phi(P_V, P_T)}, \quad R_D(G) = \frac{1}{|\mathcal{F}|} \sum_{a \in \mathcal{F}} \frac{1}{|\Phi_a|} \sum_{\phi \in \Phi_a} r_{\phi,D}(G). \tag{18}$$

We average readouts within each family before weighting the six families equally. The five discrepancies are FD, sliced W_2^2 (SWD), RBF-MMD, and sliced estimates of both KL directions; $R_{\text{avg}5} = \frac{1}{5}\sum_D R_D$. The reference ratio 1 is empirical, not a theoretical lower bound. Appendix L.3 specifies the common estimators; Appendix L.2 distinguishes this recomputed panel from published FID values.

Table 7: Family-balanced multi-representation distribution matching on ImageNet-256 (lower is better). Readouts (e.g., CLS and patch-averaged pooling) are first averaged within each encoder family, and the six encoder families are then weighted equally. The upper block compares directly trained one-step generators. AMFD-U is a multi-representation post-trained system and is listed separately; because AMFD-U trains on SigLIP2, Inception, and MAE (S+I+M), its common-panel numbers are not fully held out.

Method	$R_{\mathrm{FD}} \downarrow$	$R_{\mathrm{SWD}} \downarrow$	$R_{\mathrm{MMD}} \downarrow$	$R_{\mathrm{KL}(P_T \Vert P_G)} \downarrow$	$R_{\mathrm{KL}(P_G \Vert P_T)} \downarrow$	$R_{\mathrm{avg5}} \downarrow$
W-Flow-B	7.983	10.143	12.114	1.602	1.326	6.634
Drifting-L	7.173	9.006	11.173	1.568	1.282	6.040
W-Flow-L	6.729	8.237	9.821	1.486	1.252	5.505
pMF-L	5.954	6.758	7.997	1.402	1.189	4.660
GFD-B (ours)	5.520	4.554	5.322	1.244	1.132	3.555
GFD-L (ours)	**3.882**	**3.307**	**3.750**	**1.164**	**1.075**	**2.635**
Multi-representation post-trained baselines:						
AMFD-U (pMF-L)	1.518	1.529	1.694	1.021	1.031	1.359

GFD-L has the lowest family-balanced value for each discrepancy among the directly trained models, with $R_{\mathrm{avg5}} = 2.635$. This aggregate result is representation dependent: Drifting-L has lower FD ratios in ConvNeXtV2, Inception, and DreamSim. Nevertheless, recomputing the aggregate after removing each family in turn leaves GFD-L first among direct models in all six cases (Appendix L.5). Its aggregate lead therefore does not depend on any single family, without implying a universal ordering across representations or aggregation rules.

AMFD-U attains an aggregate of 1.359 after post-training pMF-L with SigLIP2, Inception, and MAE (S+I+M). SigLIP2 and Inception overlap the evaluation panel, whereas MAE is an additional training representation. We report it as a post-trained system-level reference.

5 Limitations

The analysis concerns classical population flows. Long-time convergence and the effect of finite neural parameter updates remain open. The reference-sample bias expansion is a fixed-kernel result under moment and denominator conditions; joint limits with increasing concentration or vanishing bandwidth require further analysis. Exact energy-mixture identities concern population fields before numerical scale balancing; balancing and finite targets follow drifting-style implementation conventions. With non-injective image encoders, distributional identification concerns the induced feature law rather than the full image law.

6 Conclusion

GFD establishes a unified, directly estimable energy-defined framework for one-step generative modeling. The clean-measure construction maps general KDE f-divergences, the quadratic KDE energy, geometry-dependent kernels, and compatible compositions to explicit particle fields and a common estimator-to-generator interface. GFD leads the controlled CIFAR-10 comparison and the family-balanced ImageNet evaluation among the compared directly trained models. The five-seed study demonstrates the value of composition for mean quality and seed stability, while encoder-free geometric and high-dimensional sequence tasks establish the framework's modeling range.

AI USE STATEMENT

In this work, we used generative AI tools for providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, implementing methods, assisting with translation, providing feedback on research methodology and experiments, and interpreting results. We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, cleaning or reformatting datasets, or supporting qualitative and thematic data analysis. No additional required-disclosure tasks are designated as not applicable to this work. Additionally, we used generative AI tools for creating or editing software code, drafting parts of the research paper, sourcing/searching for information, summarizing and analyzing existing literature, and editing the paper to improve readability. We have reviewed all AI-assisted work. Specifically, we manually verified mathematical derivations and proofs, reviewed and tested code for correctness and checked them against experimental records, and verified citations and related claims against the original sources. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The population assumptions and proofs are given in Appendices A and B. Kernel instances, field estimators, and stopped-target amortization are detailed in the appendices. Appendix K describes the experimental protocols, including multi-level representation, grouping, and kernel-temperature components; Appendix L specifies the multi-representation evaluation and reports the individual readouts and aggregation sensitivity. Reported baseline FID values and independently recomputed checkpoint evaluations are identified separately.

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. *arXiv preprint arXiv:1906.04370*, 2019.

Arwen Bradley and Preetum Nakkiran. Classifier-free guidance is a predictor-corrector. *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2408.09000.

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

Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.

Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

Imre Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

Oscar Davis, Samuel Kessler, Mircea Petrache, İsmail İ Ceylan, Michael Bronstein, and Avishek J Bose. Fisher flow matching for generative modeling over discrete data. *Advances in Neural Information Processing Systems*, 37:139054–139084, 2024.

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

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. URL https://arxiv.org/abs/2605.10727.

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

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. URL https://arxiv.org/abs/2306.09344.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1608–1617, 2018. URL https://proceedings.mlr.press/v84/genevay18a.html.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *Advances in Neural Information Processing Systems*, 2025.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Arthur Gretton, Li Kevin Wenliang, Alexandre Galashov, James Thornton, Valentin De Bortoli, and Arnaud Doucet. On the wasserstein gradient flow interpretation of drifting models. *arXiv preprint arXiv:2605.05118*, 2026.

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*, 2026a.

Jiaqi Han, Puheng Li, Qiushan Guo, Renyuan Xu, Stefano Ermon, and Emmanuel J. Candès. W-Flow: Official implementation, drift_loss_ot.py. https://github.com/hanjq17/W-Flow/blob/main/drift_loss_ot.py, 2026b.

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. In *Advances in Neural Information Processing Systems*, 2020.

Peter Holderrieth, Marton Havasi, Jason Yim, Neta Shaul, Itai Gat, Tommi Jaakkola, Brian Karrer, Ricky T. Q. Chen, and Yaron Lipman. Generator matching: Generative modeling with arbitrary markov processes. In *International Conference on Learning Representations*, 2025.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2401.09603.

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.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In *International Conference on Learning Representations*, 2024.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fréchet inception distance. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=4oXTQ6m_ws8.

Chieh-Hsin Lai, Bac Nguyen, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yuki Mitsufuji, Stefano Ermon, and Molei Tao. A unified view of score-based and drifting models. *arXiv preprint arXiv:2603.07514*, 2026.

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. URL https://arxiv.org/abs/1705.08584.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Song Liu, Jiahao Yu, Jack Simons, Mingxuan Yi, and Mark Beaumont. Minimizing f-divergences by interpolating velocity fields. *arXiv preprint arXiv:2305.15577*, 2023.

Wenze Liu, Xintao Wang, Pengfei Wan, and Xiangyu Yue. Amortized moment matching for visual generation. *arXiv preprint arXiv:2607.26860*, 2026. URL https://arxiv.org/abs/2607.26860.

Yunpeng Liu, Boxiao Liu, Yi Zhang, Xingzhong Hou, Guanglu Song, Yu Liu, and Haihang You. See further when clear: Curriculum consistency model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18103–18112, 2025.

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, 2026. URL https://arxiv.org/abs/2601.22158.

Emile Mathieu and Maximilian Nickel. Riemannian continuous normalizing flows. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2503–2515, 2020.

Axel Sauer, Katja Schwarz, and Andreas Geiger. StyleGAN-XL: Scaling stylegan to large diverse datasets. *ACM Transactions on Graphics*, 41(4), 2022.

SciPy contributors. Generalized normal distribution. SciPy reference documentation, 2026a. URL https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.gennorm.html.

SciPy contributors. Logistic distribution. SciPy reference documentation, 2026b. URL https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.logistic.html.

SciPy contributors. Multivariate student-t distribution. SciPy reference documentation, 2026c. URL https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.multivariate_t.html.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. *arXiv preprint arXiv:2310.14189*, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 3732–3784, 2023. URL https://arxiv.org/abs/2306.04675.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

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. URL https://arxiv.org/abs/2502.14786.

Erkan Turan, Nicolas Dufour, and Maks Ovsjanikov. Generative drifting is secretly score matching: a spectral and variational perspective. *arXiv preprint arXiv:2603.09936*, 2026.

Chenguang Wang and Tianshu Yu. A unified framework for data-free one-step sampling via wasserstein gradient flows. *arXiv preprint arXiv:2605.17808*, 2026.

Yifei Wang, Weimin Bai, Colin Zhang, Debing Zhang, Weijian Luo, and He Sun. Uni-instruct: One-step diffusion model through unified diffusion divergence instruction. In *Advances in Neural Information Processing Systems*, 2025.

Romann M. Weber. The score-difference flow for implicit generative modeling. *arXiv preprint arXiv:2304.12906*, 2023.

Dongyeop Woo, Marta Skreta, Seonghyun Park, Kirill Neklyudov, and Sungsoo Ahn. Riemannian meanflow, 2026. URL https://arxiv.org/abs/2602.07744.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023.

Andrew T. A. Wood. Simulation of the von Mises Fisher distribution. *Communications in Statistics - Simulation and Computation*, 23(1):157–164, 1994.

Yilun Xu, Ziming Liu, Yonglong Tian, Shangyuan Tong, Max Tegmark, and Tommi Jaakkola. PFGM++: Unlocking the potential of physics-inspired generative models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38566–38591, 2023.

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

Lihe Yang, Shang-Wen Li, Yang Li, Xinjie Lei, Dong Wang, Abdelrahman Mohamed, Hengshuang Zhao, and Hu Xu. In pursuit of pixel supervision for visual pre-training. *arXiv preprint arXiv:2512.15715*, 2025. URL https://arxiv.org/abs/2512.15715.

Jiaru Zhang, Zeyun Deng, Juanwu Lu, Ziran Wang, and Ruqi Zhang. Analytical correction for subsampling bias in drifting models. *arXiv preprint arXiv:2604.27239*, 2026.

Bowen Zheng and Tianming Yang. Revisiting diffusion models: From generative pre-training to one-step generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 78434–78453, 2025.

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 of *Proceedings of Machine Learning Research*, pp. 62307–62331, 2024.

Mingyuan Zhou, Huangjie Zheng, Yi Gu, Zhendong Wang, and Hai Huang. Adversarial score identity distillation: Rapidly surpassing the teacher in one step. In *International Conference on Learning Representations*, 2025.

Chenchen Zhu, Saksham Suri, Cijo Jose, Maxime Oquab, Marc Szafraniec, Wei Wen, Yunyang Xiong, Patrick Labatut, Piotr Bojanowski, Raghuraman Krishnamoorthi, et al. Efficient universal perception encoder. *arXiv preprint arXiv:2603.22387*, 2026.

A Notation, Admissibility, and Scope

The main theorem is stated for any setting in which the displayed operations are valid. A convenient sufficient setting is: M is a connected closed Riemannian manifold; φ is strictly positive, normalized, and C^2 in both variables; p, q are probability measures; and $f \in C^2$ on the compact range of s_q/s_p. Compactness then provides positive upper and lower bounds for the KDEs and the domination required to exchange differentiation and integration.

For Proposition C.1, require $\varphi(x, \cdot) \in L^2$, integrability of the induced kernel k_φ against the relevant measures, and domination of its first derivatives. These conditions justify the quadratic KDE identity, first variation, and particle field. Density-ratio bounds are unnecessary for this energy.

On $\mathbb{R}^d$, impose the corresponding local smoothness, integrability, domination, and vanishing-boundary conditions. These depend on the tails of p, q, the kernel derivatives, and the growth of f' near zero and infinity. The dissipation identity holds along classical solutions of the stated continuity equation.

Injectivity of $\mathcal{S}_k$ gives distributional identification for both energy families: equal KDEs imply equal clean measures. For the induced quadratic kernel, this is equivalent to the characteristic-kernel property on the same measure class (Sriperumbudur et al., 2010; 2011). Appendix E.1 provides an expanded kernel and query-noise catalog, with injectivity proofs and a concrete Euclidean admissibility regime in Appendix E.10.

For later use, define the kernel score with respect to the clean particle by

$$\mathbf{b}_\varphi(\mathbf{x}, \mathbf{y}) := \nabla_{\mathbf{x}} \log \varphi(\mathbf{x}, \mathbf{y}). \tag{19}$$

Whenever $Y \sim \varphi(\mathbf{x}, \cdot)$, Eq. (6) becomes

$$\mathbf{u}_f[q](\mathbf{x}) = -\mathbb{E}\left[f'\left(\frac{s_q(Y)}{s_p(Y)} \right) \mathbf{b}_\varphi(\mathbf{x}, Y) \right]. \tag{20}$$

This identity is the common starting point for the concrete f-divergence constructions below; the MMD construction instead differentiates the RKHS mean-embedding distance as proved in Appendix C.

B Proof of the Unified KDE f-Divergence Flow

B.1 First variation

Fix any $r \in \mathcal{P}(M)$ and consider the admissible probability-mixture path $q_\varepsilon = (1-\varepsilon)q + \varepsilon r$ for $\varepsilon \in [0, 1]$. Its zero-mass direction is $\eta = r - q$. By the linearity of $\mathcal{S}_k$,

$$s_{q_\varepsilon} = s_q + \varepsilon \mathcal{S}_k(r - q). \tag{21}$$

Differentiating the energy at zero gives

$$\left. \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \right|_0 \mathcal{G}_f(q_\varepsilon; p) = \int_M f'\left(\frac{s_q(\mathbf{y})}{s_p(\mathbf{y})} \right) (\mathcal{S}_k(r-q))(\mathbf{y}) \, \mathrm{dvol}_g(\mathbf{y}) \tag{22}$$

$$= \int_M \int_M f'\left(\frac{s_q(\mathbf{y})}{s_p(\mathbf{y})} \right) \varphi(\mathbf{x}, \mathbf{y}) \, \mathrm{dvol}_g(\mathbf{y}) \, \mathrm{d}(r-q)(\mathbf{x}) \tag{23}$$

$$= \int_M \mathcal{A}_k\left[f'\left(\frac{s_q}{s_p} \right) \right](\mathbf{x}) \, \mathrm{d}(r-q)(\mathbf{x}). \tag{24}$$

Since $r \in \mathcal{P}(M)$ is arbitrary, this identifies the first variation along all probability-mixture directions. Thus, up to an additive constant,

$$\frac{\delta \mathcal{G}_f}{\delta q} = \mathcal{A}_k\left[f'\left(\frac{s_q}{s_p} \right) \right]. \tag{25}$$

Taking the negative Riemannian gradient yields Eq. (6).

B.2 Classical-flow dissipation

Let q_t satisfy $\partial_t q_t + \operatorname{div}(q_t \mathbf{u}_f[q_t]) = 0$. Then

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{G}_f(q_t; p) = \int_M \frac{\delta \mathcal{G}_f}{\delta q_t}\, \partial_t q_t \tag{26}$$

$$= -\int_M \frac{\delta \mathcal{G}_f}{\delta q_t} \operatorname{div}(q_t \mathbf{u}_f[q_t]) \tag{27}$$

$$= \int_M \left\langle \nabla \frac{\delta \mathcal{G}_f}{\delta q_t}, \mathbf{u}_f[q_t] \right\rangle \mathrm{d}q_t \tag{28}$$

$$= -\int_M \|\mathbf{u}_f[q_t]\|_g^2 \, \mathrm{d}q_t. \tag{29}$$

This proves Eq. (7). Strict convexity of f implies $D_f(s_q \| s_p) \geq 0$, with equality only when $s_q = s_p$ almost everywhere. If $\mathcal{S}_k$ is injective on probability measures, equality further implies $q = p$.

B.3 Identification from a globally vanishing clean-particle field

Assume M is connected, f is strictly convex, and $\mathcal{S}_k$ is injective. Put $r = s_q/s_p$ and $a = \mathcal{A}_k[f'(r)]$. If $\mathbf{u}_f[q](x) = 0$ for every $x \in M$, then $\nabla a = 0$ and connectedness makes a constant. Since $q - p$ has zero mass, adjointness gives

$$0 = \int_M a \, \mathrm{d}(q - p) = \int_M f'(r)(s_q - s_p) \, \mathrm{dvol}_g \tag{30}$$

$$= \int_M s_p (r - 1)\big(f'(r) - f'(1)\big) \, \mathrm{dvol}_g. \tag{31}$$

The last integrand is nonnegative and, by strict convexity, vanishes only at $r = 1$. Thus $s_q = s_p$, and injectivity implies $q = p$. Conversely, $q = p$ makes $f'(r) = f'(1)$ constant; normalization gives $\mathcal{A}_k 1 = 1$, so the clean-particle field is zero everywhere. This proves identification by the global field identity as well as by zero energy. The field identity here is over all optimization coordinates; vanishing only at a finite collection of particles is a different condition.

B.4 Identification through an injective feature map

Let $E : \mathcal{X} \to M$ be a measurable injection between standard Borel spaces, with measurable inverse on $E(\mathcal{X})$, and let P, Q be probability measures on $\mathcal{X}$. If $E_\# P = E_\# Q$, then for every measurable $A \subseteq \mathcal{X}$,

$$P(A) = (E_\# P)(E(A)) = (E_\# Q)(E(A)) = Q(A). \tag{32}$$

Thus equality in the optimization space identifies the original distribution whenever the feature map is injective on the relevant data domain. Combined with injective smoothing and strict convexity, zero feature-space GFD energy implies $P = Q$. A continuous injection on a compact data domain is a sufficient setting for the required inverse measurability.

C KDE-L^2: Quadratic KDE Energy and Its Particle Field

Let $\varphi(x, \cdot) \in L^2(M, \mathrm{dvol}_g)$ and assume the integrability and differentiation conditions in Assumption 3.2. Define the induced kernel

$$k_\varphi(x, x') = \langle \varphi(x, \cdot), \varphi(x', \cdot) \rangle_{L^2} = \int_M \varphi(x, y)\varphi(x', y) \, \mathrm{dvol}_g(y). \tag{33}$$

This is positive semidefinite, since $\sum_{i,j} c_i c_j k_\varphi(x_i, x_j) = \| \sum_i c_i \varphi(x_i, \cdot) \|_{L^2}^2 \geq 0$.

Proposition C.1 (Quadratic KDE identity and clean-particle flow)**.** *For the KDE-L^2 energy* $\mathcal{G}_2(q;p) = \|\mathcal{S}_k q - \mathcal{S}_k p\|_{L^2}^2$,

$$\mathcal{G}_2(q;p) = \iint k_\varphi(x,x')\,\mathrm{d}(q-p)(x)\,\mathrm{d}(q-p)(x'), \tag{34}$$

$$\frac{\delta \mathcal{G}_2}{\delta q}(x) = 2\mathcal{A}_k[s_q - s_p](x) = 2\int k_\varphi(x,x')\,\mathrm{d}(q-p)(x'), \tag{35}$$

$$\mathbf{u}_2[q](x) = 2\int \nabla_1 k_\varphi(x,x')\,\mathrm{d}(p-q)(x'). \tag{36}$$

Along a classical continuity-equation flow with this field,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{G}_2(q_t;p) = -\int_M \|\mathbf{u}_2[q_t]\|_g^2\,\mathrm{d}q_t \le 0. \tag{37}$$

If $\mathcal{S}_k$ is injective, $\mathcal{G}_2(q;p) = 0$ if and only if $q = p$.

Proof. Write $\nu = q - p$. Fubini's theorem gives

$$\|\mathcal{S}_k\nu\|_{L^2}^2 = \int_M \left[\int_M \varphi(x,y)\,\mathrm{d}\nu(x)\right]\left[\int_M \varphi(x',y)\,\mathrm{d}\nu(x')\right]\mathrm{dvol}_g(y) \tag{38}$$

$$= \iint k_\varphi(x,x')\,\mathrm{d}\nu(x)\,\mathrm{d}\nu(x'). \tag{39}$$

For any zero-mass admissible perturbation η, differentiating $\|\mathcal{S}_k\nu + \epsilon\mathcal{S}_k\eta\|_{L^2}^2$ at zero gives $2\langle\mathcal{S}_k\nu, \mathcal{S}_k\eta\rangle_{L^2} = 2\int \mathcal{A}_k[s_q - s_p]\,\mathrm{d}\eta$. This proves the first variation and its negative gradient. Substitution into the continuity equation and integration by parts gives Eq. (37). Zero energy means equal KDEs; injectivity gives equal clean measures. On connected M, a globally zero field also implies zero energy: $a = \mathcal{A}_k[s_q - s_p]$ is constant, so $\int a\,\mathrm{d}(q-p) = \|s_q - s_p\|_{L^2}^2 = 0$.

C.1 Pairwise training estimator

For independent references $X_j^+ \sim p$ and $X_j^- \sim q$, the density-ratio-free estimator is

$$\widehat{\mathbf{u}}_2(x) = 2\left[\frac{1}{n_p}\sum_{j=1}^{n_p}\nabla_1 k_\varphi(x, X_j^+) - \frac{1}{n_q}\sum_{j=1}^{n_q}\nabla_1 k_\varphi(x, X_j^-)\right]. \tag{40}$$

This estimator is unbiased at a fixed particle independent of the references. It can be combined with estimated KDE f-fields before constructing the stopped target.

C.2 Gaussian and spherical induced kernels

For $\varphi_h(x,y) = \mathcal{N}(y; x, h^2 I)$, completing the square in Eq. (33) yields

$$k_{\varphi_h}(x,x') = (4\pi h^2)^{-d/2}\exp\left(-\frac{\|x-x'\|^2}{4h^2}\right). \tag{41}$$

Thus the induced pairwise kernel has bandwidth $\sqrt{2}h$, and $\nabla_1 k_{\varphi_h} = k_{\varphi_h}(x'-x)/(2h^2)$. For a vMF smoothing kernel on $\mathbb{S}^{d-1}$,

$$k_{\varphi_\kappa}(x,x') = C_d(\kappa)^2\int_{\mathbb{S}^{d-1}} e^{\kappa(x+x')^\top y}\,\mathrm{dvol}_g(y) = \frac{C_d(\kappa)^2}{C_d(\kappa\|x+x'\|)}. \tag{42}$$

At $x' = -x$, use $C_d(0) = 1/|\mathbb{S}^{d-1}|$. Tangential differentiation or differentiation under the integral supplies the spherical pairwise field.

C.3 Connection with kernel mean discrepancies

The feature map $x \mapsto \varphi(x, \cdot)$ realizes k_φ in L^2. Hence the classical identity $\mathrm{MMD}^2_{k_\varphi}(q, p) = \mathcal{G}_2(q; p)$ follows directly, and the induced pairwise dynamics agree with the corresponding MMD gradient flow (Arbel et al., 2019). In the GFD construction, the primitive modeling choice is the quadratic KDE energy, called KDE-L^2 here. The equivalence uses the induced kernel k_φ, not an arbitrary pairwise kernel with the same bandwidth as φ; the Gaussian bandwidth conversion above is part of the identity. If $\mathcal{S}_k$ is injective, k_φ is characteristic on the same probability class. These connections identify the classical special case while expressing the quadratic component in the same density-space language as the other GFD energies.

D Concrete KDE f-Divergences

Only f' enters the particle field. Replacing $f(r)$ by $\widetilde{f}(r) = f(r) + a(r-1) + b$ gives

$$D_{\widetilde{f}}(s_q \| s_p) = D_f(s_q \| s_p) + b, \tag{43}$$

because both KDE densities integrate to one. Thus $a(r-1)$ leaves the divergence unchanged, while b shifts the energy by a constant. Neither term changes the first variation or particle field: the added constant a in $\widetilde{f}'$ vanishes after $\nabla_{\mathbf{x}} \int \varphi(\mathbf{x}, \mathbf{y}) \mathrm{dvol}_g(\mathbf{y}) = 0$. It is therefore useful to use response functions normalized so that $f'(1) = 0$.

Let

$$r(\mathbf{y}) := \frac{s_q(\mathbf{y})}{s_p(\mathbf{y})}. \tag{44}$$

The master field is

$$\mathbf{u}_f[q](\mathbf{x}) = -\mathbb{E}_{Y \sim \varphi(\mathbf{x}, \cdot)} \left[\psi_f(r(Y)) \mathbf{b}_\varphi(\mathbf{x}, Y) \right], \qquad \psi_f = f'. \tag{45}$$

D.1 Forward KDE-KL

Choose

$$f_{\mathrm{FKL}}(r) = r \log r - r + 1, \qquad \psi_{\mathrm{FKL}}(r) = \log r. \tag{46}$$

Then

$$\mathbf{u}_{\mathrm{FKL}}(\mathbf{x}) = -\mathbb{E}_Y \left[\log \frac{s_q(Y)}{s_p(Y)} \mathbf{b}_\varphi(\mathbf{x}, Y) \right]. \tag{47}$$

For a Gaussian kernel this field has the additional score-difference form proved in Appendix E.11.

D.2 Reverse KDE-KL

Using the orientation $D_f(s_q \| s_p)$, reverse KL is obtained with

$$f_{\mathrm{RKL}}(r) = -\log r + r - 1, \qquad \psi_{\mathrm{RKL}}(r) = 1 - r^{-1}. \tag{48}$$

The constant 1 disappears after kernel averaging, so either of the equivalent forms

$$\mathbf{u}_{\mathrm{RKL}}(\mathbf{x}) = -\mathbb{E}_Y \left[(1 - r(Y)^{-1}) \mathbf{b}_\varphi(\mathbf{x}, Y) \right] \tag{49}$$

$$= -\mathbb{E}_Y \left[-r(Y)^{-1} \mathbf{b}_\varphi(\mathbf{x}, Y) \right] \tag{50}$$

may be used.

D.3 Pearson χ^2 and squared Hellinger

For the half-scaled Pearson divergence,

$$f_{\chi^2}(r) = \frac{1}{2}(r-1)^2, \qquad \psi_{\chi^2}(r) = r - 1, \tag{51}$$

so

$$\mathbf{u}_{\chi^2}(\mathbf{x}) = -\mathbb{E}_Y \left[(r(Y) - 1) \mathbf{b}_\varphi(\mathbf{x}, Y) \right]. \tag{52}$$

For squared Hellinger,

$$f_{\mathrm{Hel}}(r) = 2(\sqrt{r}-1)^2, \qquad \psi_{\mathrm{Hel}}(r) = 2(1-r^{-1/2}), \tag{53}$$

and hence

$$\mathbf{u}_{\mathrm{Hel}}(\mathbf{x}) = -2\mathbb{E}_Y\left[(1-r(Y)^{-1/2})\mathbf{b}_\varphi(\mathbf{x},Y)\right]. \tag{54}$$

A global positive rescaling of f only rescales flow time and can be absorbed into the particle step size.

D.4 The α-divergence family

For $\alpha \in \mathbb{R}\setminus\{0,1\}$, define

$$f_\alpha(r) = \frac{r^\alpha - \alpha r + \alpha - 1}{\alpha(\alpha-1)}, \qquad \psi_\alpha(r) = \frac{r^{\alpha-1}-1}{\alpha-1}. \tag{55}$$

The limits $\alpha \to 1$ and $\alpha \to 0$ recover forward and reverse KL, respectively, while $\alpha = 2$ gives the half-scaled Pearson divergence. The corresponding field is

$$\mathbf{u}_\alpha(\mathbf{x}) = -\mathbb{E}_Y\left[\frac{r(Y)^{\alpha-1}-1}{\alpha-1}\mathbf{b}_\varphi(\mathbf{x},Y)\right]. \tag{56}$$

This family provides a continuous interpolation of density-ratio responses without changing the GFD derivation.

E Kernel Instantiations and Explicit Particle Fields

E.1 Selecting a kernel and its query-noise distribution

Choose the optimization manifold and a positive interaction function $K(x,y)$ with finite, differentiable normalizer $Z(x) = \int_M K(x,y)\,\mathrm{dvol}_g(y)$. The normalized kernel and its clean-particle score are

$$\varphi(x,y) = \frac{K(x,y)}{Z(x)}, \qquad b_\varphi(x,y) = \nabla_x \log K(x,y) - \nabla_x \log Z(x). \tag{57}$$

Kernel selection therefore fixes both the interaction profile and the conditional query law $Y \mid x \sim \varphi(x,\cdot)\,\mathrm{dvol}_g$. A sampler must realize this normalized law. Translation-invariant kernels yield additive noise; kernels on manifolds yield a conditional distribution on the manifold. If Z depends on x, its derivative is part of the field. For distributional identification, choose an injective smoothing operator; for KDE-L^2, also require square-integrable kernel sections. Bandwidth or concentration controls the scale of interactions and is held fixed when evaluating each energy component.

An admissible kernel catalog. Table 8 summarizes normalized kernels whose query laws can be sampled explicitly. All listed kernels are strictly positive, C^2, and have square-integrable sections and injective smoothing operators under the displayed parameter restrictions. The following subsections derive their normalization, sampling laws, and clean-particle scores. A concrete sufficient regime for the Euclidean entries is compactly supported clean measures and the KL, reverse-KL, Pearson, or Hellinger responses in Table 1; the tail argument is given below. On the sphere and torus, compactness supplies the admissibility bounds for any $f \in C^2$ on the attained density-ratio range.

Gaussian selection and additive noise. On $\mathbb{R}^d$, choose $K_h(x,y) = \exp(-\|y-x\|^2/(2h^2))$ with $h > 0$. The substitution $z = (y-x)/h$ gives $Z = (2\pi h^2)^{d/2}$, independent of x. Thus $Y = x + h\epsilon$, $\epsilon \sim \mathcal{N}(0,I)$, samples the kernel exactly, and $b_\varphi = (Y-x)/h^2$. The parameter h is the standard deviation of the query perturbation in each coordinate. Substituting this score in Eq. (20) gives the Gaussian field below.

Table 8: Kernel choices and their induced query noise. Each sampler generates the normalized conditional law $\varphi(x, \cdot)$. Parameters are fixed within an energy component.

Kernel	Space and parameters	Query-noise construction
Gaussian	$\mathbb{R}^d$, $\Sigma \succ 0$	$Y = x + L\epsilon$, $LL^\top = \Sigma$, $\epsilon \sim \mathcal{N}(0, I)$
Radial generalized Gaussian	$\mathbb{R}^d$, $h > 0$, $\beta \geq 2$	Gamma radius and independent uniform direction
Student-t	$\mathbb{R}^d$, $\Sigma \succ 0$, $\nu > 2$	Gaussian noise scaled by an independent inverse chi-square root
Product Logistic	$\mathbb{R}^d$, scales $a_i > 0$	Independent scaled logits of uniform variables
Gaussian mixture	$\mathbb{R}^d$, $\pi_j > 0$, $\Sigma_j \succ 0$	Categorical component followed by Gaussian noise
Wrapped Gaussian	$\mathbb{T}^d$, $\Sigma \succ 0$	Gaussian displacement reduced modulo 2π
von Mises–Fisher	$\mathbb{S}^{d-1}$, $\kappa > 0$	Directional vMF query via Wood sampling

vMF selection and directional noise. On $\mathbb{S}^{d-1}$, choose $K_\kappa(x, y) = e^{\kappa x^\top y}$ with $\kappa > 0$. Rotational invariance makes the normalizer independent of x, and spherical integration gives $Z = (2\pi)^{d/2} I_{d/2-1}(\kappa)/\kappa^{d/2-1}$. Hence the query distribution is $\mathrm{vMF}(x, \kappa)$. The parameter κ controls angular concentration; projecting the ambient derivative κy onto $T_x\mathbb{S}^{d-1}$ gives its intrinsic kernel score. The derivation below yields the spherical field from the same expectation identity.

An exact vMF query can be generated using Wood's rejection sampler (Wood, 1994). For $d \geq 2$, put $m = d - 1$ and

$$b = \frac{m}{2\kappa + \sqrt{4\kappa^2 + m^2}}, \quad x_0 = \frac{1 - b}{1 + b}, \quad c = \kappa x_0 + m \log(1 - x_0^2). \tag{58}$$

Draw $Z \sim \mathrm{Beta}(m/2, m/2)$ and $U \sim \mathrm{Uniform}(0, 1)$ independently, set $W = [1 - (1 + b)Z]/[1 - (1 - b)Z]$, and accept when $\kappa W + m \log(1 - x_0 W) - c \geq \log U$. Repeat until acceptance. Independently draw V uniformly on $\mathbb{S}^{d-2}$ and choose an orthonormal tangent basis $B_x \in \mathbb{R}^{d \times (d-1)}$. Then

$$Y = Wx + \sqrt{1 - W^2}\, B_x V \tag{59}$$

has the required vMF law. At $\kappa = 0$, sample uniformly on the sphere. This construction specifies directional noise directly through the normalized spherical kernel.

Injectivity of the two instances. For Gaussian smoothing, the Fourier multiplier is $\exp(-h^2\|\omega\|^2/2) > 0$. Equality of smoothed measures therefore implies equality of their Fourier transforms and hence of the measures. For vMF smoothing with $\kappa > 0$, let $\nu = q - p$ and suppose $\mathcal{S}_k \nu = 0$. Uniform convergence of the exponential series on the compact sphere gives

$$0 = \iint e^{\kappa x^\top y}\, \mathrm{d}\nu(x)\, \mathrm{d}\nu(y) = \sum_{j=0}^{\infty} \frac{\kappa^j}{j!} \left\| \int x^{\otimes j}\, \mathrm{d}\nu(x) \right\|^2. \tag{60}$$

Every polynomial moment of ν vanishes. Polynomials restricted to the sphere are dense in continuous functions, so $\nu = 0$. Thus both instances supply the injective smoothing required by Theorem 3.3; the uniform kernel at $\kappa = 0$ does not.

Product kernels and independent query components. On a product space, use $\varphi(x, y) = \prod_{r=1}^{R} \varphi_r(x_r, y_r)$. Normalization follows from Fubini, the component score is $\nabla_{x_r} \log \varphi_r(x_r, y_r)$, and the query law is generated by conditionally independent draws $Y_r \mid x_r \sim \varphi_r(x_r, \cdot)$. Joint KDEs average these products over joint reference samples, preserving dependencies among coordinates. Marginal and joint energies can then be combined by Proposition 3.6.

E.2 Translation-invariant Euclidean kernels

Let

$$\varphi_h(\mathbf{x}, \mathbf{y}) = h^{-d} \rho\left(\frac{\mathbf{y} - \mathbf{x}}{h}\right), \qquad \int_{\mathbb{R}^d} \rho(\mathbf{z})\, \mathrm{d}\mathbf{z} = 1, \tag{61}$$

where ρ is positive and differentiable. With $\Xi = (Y - \mathbf{x})/h \sim \rho$ and $\mathbf{b}_\rho(\mathbf{z}) := -\nabla_{\mathbf{z}} \log \rho(\mathbf{z})$,

$$\nabla_{\mathbf{x}} \log \varphi_h(\mathbf{x}, Y) = \frac{1}{h} \mathbf{b}_\rho(\Xi). \tag{62}$$

Therefore the general Euclidean field is

$$\boxed{\mathbf{u}_f(\mathbf{x}) = -\frac{1}{h} \mathbb{E}_{\Xi \sim \rho} \left[f'(r(\mathbf{x} + h\Xi))\, \mathbf{b}_\rho(\Xi) \right].} \tag{63}$$

E.3 Gaussian kernel

For $\rho(\mathbf{z}) = (2\pi)^{-d/2} \exp(-\|\mathbf{z}\|^2/2)$, $\mathbf{b}_\rho(\mathbf{z}) = \mathbf{z}$. Hence

$$\mathbf{u}_f(\mathbf{x}) = -\mathbb{E}_{Y \sim \mathcal{N}(\mathbf{x}, h^2 I)} \left[f'(r(Y)) \frac{Y - \mathbf{x}}{h^2} \right], \tag{64}$$

which is Corollary 3.4.

Given reference samples $\{x_j^+\}_{j=1}^{B_p}$, the Gaussian KDE and its score at a query y are

$$\widehat{s}_p(y) = \frac{1}{B_p} \sum_{j=1}^{B_p} \rho_h(y - x_j^+), \tag{65}$$

$$\nabla_y \log \widehat{s}_p(y) = \frac{\sum_j \rho_h(y - x_j^+)(x_j^+ - y)}{h^2 \sum_j \rho_h(y - x_j^+)}. \tag{66}$$

The same formula applies to model references. Thus the Gaussian KDE-KL query field is the difference of two normalized kernel centroids divided by h^2, recovering the classical mean-shift identity (Cheng, 1995; Comaniciu & Meer, 2002).

E.4 Anisotropic Gaussian kernels

Let $\Sigma = LL^\top \succ 0$ be fixed and write $z = y - x$. A change of variables $z = L\epsilon$ gives the normalized kernel

$$\varphi_\Sigma(x, y) = \frac{\exp(-z^\top \Sigma^{-1} z/2)}{(2\pi)^{d/2} |\Sigma|^{1/2}}, \qquad Y = x + L\epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{67}$$

Thus the noise covariance and the interaction anisotropy are the same modeling choice. Differentiation yields $b_\varphi(x, y) = \Sigma^{-1}(y - x)$ and

$$\mathbf{u}_f(x) = -\mathbb{E}_\epsilon \left[f'(r(x + L\epsilon)) L^{-\top} \epsilon \right]. \tag{68}$$

The Fourier multiplier $e^{-\omega^\top \Sigma \omega/2}$ is strictly positive, proving injectivity. Taking $\Sigma = h^2 I$ recovers the isotropic Gaussian instance.

E.5 Generalized-Gaussian radial kernels

For $\beta \geq 2$, choose $\rho_\beta(z) = C_{d,\beta} \exp(-\|z\|^\beta/\beta)$. Let $\omega_{d-1} = |\mathbb{S}^{d-1}|$. Polar integration and the substitution $t = r^\beta/\beta$ give

$$\int_{\mathbb{R}^d} e^{-\|z\|^\beta/\beta} \, \mathrm{d}z = \omega_{d-1} \int_0^\infty r^{d-1} e^{-r^\beta/\beta} \, \mathrm{d}r = \omega_{d-1} \beta^{d/\beta - 1} \Gamma(d/\beta). \tag{69}$$

Consequently $C_{d,\beta} = \beta^{1-d/\beta}/[\omega_{d-1}\Gamma(d/\beta)]$. The same change of variables identifies the radial noise law:

$$T \sim \mathrm{Gamma}(d/\beta, \text{rate} = 1), \quad U \sim \mathrm{Uniform}(\mathbb{S}^{d-1}), \quad Y = x + h(\beta T)^{1/\beta} U, \tag{70}$$

with T, U independent. In one dimension this is the generalized normal family, with scale adjusted by $\beta^{1/\beta}$ (SciPy contributors, 2026a); in higher dimensions the shared radius couples the coordinates. Differentiation gives $b_\varphi(x, y) = \|y - x\|^{\beta-2}(y - x)/h^\beta$, so

$$\mathbf{u}_f(x) = -\mathbb{E}_Y \left[f'(r(Y)) \frac{\|Y - x\|^{\beta-2}(Y - x)}{h^\beta} \right]. \tag{71}$$

The restriction $\beta \geq 2$ gives C^2 regularity at the kernel center; $\beta = 2$ yields Gaussian noise. For injectivity, the super-exponential tail makes the Fourier transform real analytic, and its value at zero is one. Its zero set therefore has empty interior. If $\rho_\beta * (q - p) = 0$, the continuous Fourier transform of $q - p$ vanishes wherever $\widehat{\rho}_\beta \neq 0$, a dense set, and hence vanishes everywhere. This proves injectivity even when the smoothing kernel itself is not positive definite; its induced quadratic kernel remains positive semidefinite by Eq. (33).

E.6 Student-t kernels and heavy-tailed query noise

Let $\Sigma = LL^\top \succ 0$, $\nu > 2$, and $a = z^\top \Sigma^{-1} z$ for $z = y - x$. The normalized Student-t kernel (SciPy contributors, 2026c) is

$$\varphi_{\nu,\Sigma}(x, y) = \frac{\Gamma((\nu + d)/2)}{\Gamma(\nu/2)(\nu\pi)^{d/2}|\Sigma|^{1/2}} \left(1 + \frac{a}{\nu}\right)^{-(\nu+d)/2}. \tag{72}$$

To derive its query law, draw $V \sim \chi^2_\nu$ independently of $\epsilon \sim \mathcal{N}(0, I)$ and set $Y = x + \sqrt{\nu/V}\, L\epsilon$. Conditional on $V = v$, the noise is Gaussian with covariance $\nu\Sigma/v$. Integrating its density against $v^{\nu/2-1}e^{-v/2}/[2^{\nu/2}\Gamma(\nu/2)]$ uses the Gamma integral and yields Eq. (72). Its covariance is $\nu\Sigma/(\nu-2)$, while its clean-particle score and field are

$$b_\varphi(x, y) = \frac{\nu + d}{\nu + a}\Sigma^{-1}(y - x), \qquad \mathbf{u}_f(x) = -\mathbb{E}_Y \left[f'(r(Y)) \frac{\nu + d}{\nu + a(Y)} \Sigma^{-1}(Y - x) \right]. \tag{73}$$

This kernel gives heavy-tailed queries with bounded displacement scores. Its Fourier multiplier is $\mathbb{E}_V \exp[-\nu\omega^\top \Sigma\omega/(2V)] > 0$, so smoothing is injective. The finite-variance choice $\nu > 2$ supplies a simple parameter regime for query sampling.

E.7 Product Logistic kernels and inverse-CDF noise

For coordinate scales $a_i > 0$, define

$$\varphi_{\boldsymbol{a}}(x, y) = \prod_{i=1}^{d} \frac{1}{4a_i} \operatorname{sech}^2\left(\frac{y_i - x_i}{2a_i}\right). \tag{74}$$

Each factor is the derivative of the Logistic CDF $F_i(z) = (1+e^{-z/a_i})^{-1}$ (SciPy contributors, 2026b). Its integral is one, and inversion gives the exact noise construction

$$U_i \overset{\text{iid}}{\sim} \text{Uniform}(0, 1), \qquad Y_i = x_i + a_i \log \frac{U_i}{1 - U_i}. \tag{75}$$

Logarithmic differentiation gives $[b_\varphi(x, y)]_i = a_i^{-1} \tanh[(y_i - x_i)/(2a_i)]$, hence

$$[\mathbf{u}_f(x)]_i = -\mathbb{E}_Y \left[f'(r(Y)) \frac{1}{a_i} \tanh\left(\frac{Y_i - x_i}{2a_i}\right) \right]. \tag{76}$$

The noise has exponential tails and coordinate variances $\pi^2 a_i^2/3$. With $t = a_i\omega_i$, the inverse-CDF substitution gives the characteristic function $\int_0^1 u^{it}(1 - u)^{-it}\,\mathrm{d}u = \Gamma(1 + it)\Gamma(1 - it) = \pi t/\sinh(\pi t) > 0$, using its continuous value one at $t = 0$. The product multiplier is therefore nonzero, proving injectivity.

E.8 Gaussian-mixture kernels and categorical query noise

Choose fixed weights $\pi_j > 0$, $\sum_{j=1}^J \pi_j = 1$, and covariances $\Sigma_j = L_j L_j^\top \succ 0$. The centered Gaussian-mixture kernel is

$$\varphi(x, y) = \sum_{j=1}^{J} \pi_j \varphi_{\Sigma_j}(x, y). \tag{77}$$

Normalization follows by summing unit integrals. Sampling $C \sim \text{Categorical}(\pi)$ followed by $Y = x + L_C\epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, realizes exactly this kernel. Define the posterior component weights

$w_j(x,y) = \pi_j \varphi_{\Sigma_j}(x,y)/\varphi(x,y)$. The score and field become

$$b_\varphi(x,y) = \sum_j w_j(x,y)\Sigma_j^{-1}(y-x), \tag{78}$$

$$\mathbf{u}_f(x) = -\sum_j \pi_j \mathbb{E}_{Y\sim\varphi_{\Sigma_j}(x,\cdot)}\left[f'(r_{\mathrm{mix}}(Y))\Sigma_j^{-1}(Y-x)\right]. \tag{79}$$

Here r_{mix} is computed from KDEs using the full mixture kernel. This provides multiple interaction scales within one normalized smoothing operator. Mixing kernels and mixing energies are distinct constructions: in the former, all query components share r_{mix}; in the latter, separate energy responses are combined as in Proposition 3.6. The Fourier multiplier $\sum_j \pi_j e^{-\omega^\top \Sigma_j \omega/2} > 0$ establishes injectivity.

E.9 Wrapped Gaussian kernels and periodic query noise

For periodic coordinates, let $M = \mathbb{T}^d = \mathbb{R}^d/(2\pi\mathbb{Z})^d$ with its flat metric. With $\Sigma = LL^\top \succ 0$, define

$$\varphi_\Sigma^{\mathrm{wrap}}(x,y) = \sum_{m\in\mathbb{Z}^d} \rho_\Sigma(y-x+2\pi m), \qquad \rho_\Sigma(z) = \mathcal{N}(z;0,\Sigma). \tag{80}$$

Integrating over a fundamental cell tiles $\mathbb{R}^d$, so the kernel has unit integral. Gaussian decay permits termwise differentiation, and the wrapped displacement

$$Y = (x + L\epsilon) \bmod 2\pi, \qquad \epsilon \sim \mathcal{N}(0,I), \tag{81}$$

has this density because all integer lifts of Y contribute to the sum. Put $z_m = y - x + 2\pi m$. Then

$$b_\varphi(x,y) = \frac{\sum_m \rho_\Sigma(z_m)\Sigma^{-1}z_m}{\sum_m \rho_\Sigma(z_m)}, \qquad \mathbf{u}_f(x) = -\mathbb{E}_Y[f'(r(Y))b_\varphi(x,Y)]. \tag{82}$$

The noise law is sampled by one wrapped Gaussian draw; evaluating the score averages the possible windings. Each Fourier mode $k \in \mathbb{Z}^d$ has smoothing multiplier $e^{-k^\top \Sigma k/2} > 0$, so the kernel identifies probability measures on the torus. Positivity, smoothness, and compactness provide the remaining admissibility bounds.

E.10 Sufficient conditions for the Euclidean catalog

Every Euclidean density above is bounded, normalized, and strictly positive, so its sections are square-integrable. Its derivatives up to second order are integrable; Gaussian and generalized-Gaussian tails absorb their polynomial factors, Student-t derivatives decay faster than its density, and Logistic scores are bounded. The injectivity arguments above apply to all finite probability measures.

For a concrete energy-level admissibility regime, let the supports of the clean measures lie in a common compact set. For the Gaussian, finite Gaussian-mixture, and radial generalized-Gaussian kernels, uniformly over clean locations one has $|\log r(y)| \le C(1+\|y\|^{\beta-1})$ and kernel tails of order $e^{-c\|y\|^\beta}$, with $\beta = 2$ for the Gaussian cases. Thus polynomial powers of r or r^{-1} and logarithmic responses, multiplied by kernel derivatives, remain integrable. For Student-t and Logistic kernels, bounded displacement scores bound the density ratio above and away from zero for compact clean supports, and kernel derivatives supply integrable domination. This verifies the first-variation operations for all four responses in Table 1; classical-flow dissipation then uses the integration-by-parts condition of Assumption 3.2. Other response and tail combinations are covered by the same operation-level conditions.

E.11 Gaussian KL integration by parts

Proposition E.1 (SDF relation)**.** *Define the proxy SDF field* $v_h(\mathbf{y}) = \nabla \log s_p(\mathbf{y}) - \nabla \log s_q(\mathbf{y})$. *Then*

$$\mathbf{u}_{\mathrm{KL}}[q](\mathbf{x}) = \mathbb{E}[v_h(\mathbf{x}+\mathbf{z}) \mid \mathbf{x}]. \tag{83}$$

Consequently, $v_h(\mathbf{x}+\mathbf{z})$ *is a one-query conditionally unbiased estimator of the clean-measure KDE-KL field.*

For forward KL, the additive constant in $f'(r) = \log r + 1$ vanishes by kernel normalization, so

$$\mathbf{u}_{\mathrm{KL}}(\mathbf{x}) = -\nabla_{\mathbf{x}} \int \rho_h(\mathbf{y}-\mathbf{x}) \log \frac{s_q(\mathbf{y})}{s_p(\mathbf{y})}\, \mathrm{d}\mathbf{y}. \tag{84}$$

Since $\nabla_{\mathbf{x}} \rho_h(\mathbf{y}-\mathbf{x}) = -\nabla_{\mathbf{y}} \rho_h(\mathbf{y}-\mathbf{x})$, integration by parts gives

$$\mathbf{u}_{\mathrm{KL}}(\mathbf{x}) = -\int \rho_h(\mathbf{y}-\mathbf{x}) \nabla_{\mathbf{y}} \log \frac{s_q(\mathbf{y})}{s_p(\mathbf{y})}\, \mathrm{d}\mathbf{y} \tag{85}$$

$$= \mathbb{E}_{Y \sim \mathcal{N}(\mathbf{x}, h^2 I)} \left[\nabla \log s_p(Y) - \nabla \log s_q(Y) \right]. \tag{86}$$

This proves Eq. (11) and the identity in Eq. (83) and shows that a single noisy-query SDF update is a conditionally unbiased Monte Carlo sample of the clean-measure Gaussian KDE-KL field.

E.12 von Mises–Fisher kernel on the sphere

On $\mathbb{S}^{d-1}$, let

$$\varphi_\kappa(\mathbf{x}, \mathbf{y}) = C_d(\kappa) e^{\kappa \mathbf{x}^\top \mathbf{y}}, \qquad C_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}. \tag{87}$$

The ambient derivative is $\kappa \mathbf{y}$ and projection onto $T_{\mathbf{x}}\mathbb{S}^{d-1}$ yields

$$\nabla_{\mathbb{S},\mathbf{x}} \log \varphi_\kappa(\mathbf{x}, \mathbf{y}) = \kappa (I - \mathbf{x}\mathbf{x}^\top)\mathbf{y} = \kappa\big(\mathbf{y} - (\mathbf{x}^\top \mathbf{y})\mathbf{x}\big). \tag{88}$$

Therefore

$$\mathbf{u}_f(\mathbf{x}) = -\kappa \mathbb{E}_{Y \sim \mathrm{vMF}(\mathbf{x}, \kappa)} \left[f'(r(Y)) \big(Y - (\mathbf{x}^\top Y)\mathbf{x}\big) \right]. \tag{89}$$

For samples $x_j^+ \in \mathbb{S}^{d-1}$,

$$\widehat{s}_p(y) = \frac{C_d(\kappa)}{B_p} \sum_{j=1}^{B_p} e^{\kappa (x_j^+)^\top y}. \tag{90}$$

The normalizing constant cancels in the ratio $\widehat{s}_q(y)/\widehat{s}_p(y)$ when the same κ is used, but it remains part of the normalized query law used by the population formula.

F Conditional GFD and Query-Density Guidance

F.1 Conditional KDE f-divergence

For a condition c, replace the target and model measures by p_c, q_c. The energy $\mathcal{G}_f^c = D_f(\mathcal{S}_k q_c \| \mathcal{S}_k p_c)$ gives

$$\mathbf{u}_f^c(x) = -\mathbb{E}_{Y \sim \varphi(x,\cdot)} \left[f'\!\left(\frac{s_{q,c}(Y)}{s_{p,c}(Y)} \right) b_\varphi(x, Y) \right]. \tag{91}$$

Thus the same variational construction and sample estimator apply conditionally.

F.2 KL guidance and its energy

Define the normalized guided query density

$$\widetilde{s}_{p,w}(y) = Z_w^{-1} s_{p,c}(y)^{1+w} s_{p,\varnothing}(y)^{-w}, \qquad Z_w = \int_M s_{p,c}^{1+w} s_{p,\varnothing}^{-w}\, \mathrm{dvol}_g < \infty. \tag{92}$$

The guided energy is

$$\mathcal{G}_{\mathrm{cfg}}(q_c) = \mathrm{KL}(s_{q,c} \| \widetilde{s}_{p,w}). \tag{93}$$

Its first variation is $\mathcal{A}_k[\log(s_{q,c}/\widetilde{s}_{p,w}) + 1]$. Define both component fields with the same current model KDE: $\mathbf{u}_c = -\nabla \mathcal{A}_k \log(s_{q,c}/s_{p,c})$ and $\mathbf{u}_\varnothing = -\nabla \mathcal{A}_k \log(s_{q,c}/s_{p,\varnothing})$. Linearity and $(1+w) - w = 1$ then give

$$\mathbf{u}_{\mathrm{cfg}} = (1+w)\mathbf{u}_c - w\mathbf{u}_\varnothing. \tag{94}$$

Normalization makes $\mathcal{A}_k \log Z_w$ constant, so computing the field does not require estimating Z_w. This identity applies to the normalized-kernel construction; for Gaussian queries it also takes the score form

$$\mathbf{u}_{\mathrm{cfg}}(x) = \mathbb{E}_Z\big[(1+w)\nabla \log s_{p,c}(x+Z) - w\nabla \log s_{p,\varnothing}(x+Z) - \nabla \log s_{q,c}(x+Z)\big]. \tag{95}$$

Replacing the three scores by their empirical Gaussian KDE formulas yields

$$\widehat{\mathbf{u}}_{\mathrm{cfg}}(x) = \frac{1}{L}\sum_{\ell=1}^{L}\big[(1+w)\widehat{g}_c(x+Z_\ell) - w\widehat{g}_\varnothing(x+Z_\ell) - \widehat{g}_q(x+Z_\ell)\big], \tag{96}$$

where $Z_\ell \sim \mathcal{N}(0, h^2 I)$ and $\widehat{g} = \nabla \log \widehat{s}$ is evaluated by Eq. (66). Conditional general-f fields remain available through Eq. (91); their nonnegative energy combinations follow Appendix G.

F.3 Field guidance and target-law identification

Classifier-free guidance is also used in ODE/SDE-based generation, including flow matching and diffusion (Ho & Salimans, 2022; Lipman et al., 2023). A linear combination of conditional and unconditional fields specifies guided dynamics, but does not by itself establish that the endpoint law equals a prescribed power-tilted data distribution. For diffusion CFG, this distinction is made explicit by Bradley & Nakkiran (2025). A well-posed guided ODE or SDE still has an endpoint probability law; the issue is its identification with the intended tilt.

GFD exposes the corresponding question directly at the energy level. Equation (93) is a valid query-density KL objective, with its own clean-measure first variation and dissipation. To identify a clean target $\widetilde{p}$, one further needs $\widetilde{s}_{p,w} = \mathcal{S}_k\widetilde{p}$: taking products and powers of smoothed densities does not generally preserve the range of the smoothing operator. When such a target exists and smoothing is injective, it is unique. Thus guided-field construction and identification of a particular clean target are separate questions shared with other CFG formulations.

G Nonnegative Energy Mixtures

Let $\mathcal{G}_j$ be admissible differentiable energies on the same state space, including KDE f-divergence and quadratic KDE terms, and write $\mathbf{u}_j = -\nabla(\delta\mathcal{G}_j/\delta q)$. For fixed nonnegative weights, linearity of the first variation gives

$$\frac{\delta\mathcal{G}_{\mathrm{mix}}}{\delta q} = \sum_{j=1}^{m}\alpha_j\frac{\delta\mathcal{G}_j}{\delta q}, \qquad \mathbf{u}_{\mathrm{mix}} = \sum_{j=1}^{m}\alpha_j\mathbf{u}_j, \qquad \alpha_j \geq 0. \tag{97}$$

When the continuity equation is driven by the total field, the total energy dissipates according to the squared norm of that field:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{G}_{\mathrm{mix}}(q_t) = -\int\left\|\sum_j \alpha_j\mathbf{u}_j[q_t]\right\|^2 \mathrm{d}q_t. \tag{98}$$

This proves Proposition 3.6 and applies, in particular, to mixtures containing both kernel-density f-divergence and quadratic KDE terms.

G.1 Shared-kernel f-divergence mixtures

Let

$$\mathcal{G}_{\mathrm{mix}}(q;p) = \sum_{j=1}^{m}\alpha_j D_{f_j}(s_q\|s_p), \qquad \alpha_j \geq 0. \tag{99}$$

Linearity of the first variation gives

$$\frac{\delta\mathcal{G}_{\mathrm{mix}}}{\delta q} = \mathcal{A}_k\left[\sum_j \alpha_j f_j'(r)\right], \tag{100}$$

so the mixture is equivalently described by the combined response

$$\psi_{\mathrm{mix}}(r) = \sum_j \alpha_j f_j'(r) \tag{101}$$

and field

$$\boxed{\mathbf{u}_{\mathrm{mix}}(\mathbf{x}) = -\mathbb{E}_Y\left[\psi_{\mathrm{mix}}(r(Y))\mathbf{b}_\varphi(\mathbf{x}, Y)\right] = \sum_j \alpha_j \mathbf{u}_{f_j}(\mathbf{x}).} \tag{102}$$

For a combination of KL, reverse KL, and half-scaled Pearson energies, normalized responses give

$$\psi_{\mathrm{mix}}(r) = \alpha \log r + \beta(1 - r^{-1}) + \gamma(r - 1). \tag{103}$$

The total energy dissipates according to Eq. (98), not the weighted sum of the component squared norms. A quadratic KDE term $\lambda\mathcal{G}_2$ adds $\lambda\mathbf{u}_2$ through Eq. (9). Its induced pairwise kernel is determined by its smoothing kernel, which may be chosen independently of the smoothing kernels in the f-divergence components.

H From Particle Fields to a One-Step Generator

H.1 Euclidean stopped-target gradient

Let $x = G_\theta(\epsilon)$ and construct the detached target

$$x^+ = \mathrm{sg}[x + \eta\widehat{\mathbf{u}}(x)]. \tag{104}$$

For the scaled regression loss

$$\mathcal{L}_\eta(\theta) = \frac{1}{2\eta}\mathbb{E}_\epsilon\left\|G_\theta(\epsilon) - x^+\right\|^2, \tag{105}$$

the autodifferentiated gradient is

$$\nabla_\theta\mathcal{L}_\eta = \frac{1}{\eta}\mathbb{E}_\epsilon\left[J_\theta G_\theta(\epsilon)^\top(G_\theta(\epsilon) - x^+)\right] \tag{106}$$

$$= -\mathbb{E}_\epsilon\left[J_\theta G_\theta(\epsilon)^\top\widehat{\mathbf{u}}(G_\theta(\epsilon))\right]. \tag{107}$$

Without the factor $1/(2\eta)$, the direction is identical up to a positive scale. Thus stopped-target regression pulls the ideal particle field back through the generator Jacobian.

H.2 Manifold retraction

On a manifold, let

$$x^+ = \mathrm{sg}[\mathrm{Retr}_x(\eta\widehat{\mathbf{u}}(x))]. \tag{108}$$

For a first-order retraction, $\mathrm{Retr}_x(\eta u) = \exp_x(\eta u) + O(\eta^2)$, so a squared geodesic or locally Euclidean regression loss produces the same tangent pullback as Eq. (107) up to $O(\eta)$ terms. For comparison, a first-order retraction on the unit sphere is

$$\mathrm{Retr}_x(\eta u) = \frac{x + \eta u}{\|x + \eta u\|}, \qquad u \in T_x\mathbb{S}^{d-1}. \tag{109}$$

The spherical implementation in Appendix K.1 instead uses the exact exponential map with a clipped geodesic step. The parameterized generator generally cannot realize every independent particle displacement. The regression update therefore amortizes the population field through the model's accessible parameter tangent space; it is not an assertion that finite-step parameter optimization is identical to the infinite-dimensional Wasserstein flow.

Table 9: Gradient Flow Drifting training.

Input: generator G_θ, divergence field, step η	
1	Sample data $\{x_j^+\}$ and noise $\{\epsilon_i\}$.
2	Generate clean particles $x_i = G_\theta(\epsilon_i)$.
3	Estimate the chosen divergence field from data and model references.
4	Take a particle step along that field and stop the target gradient.
5	Update θ using $\sum_i d_M^2(G_\theta(\epsilon_i), \mathrm{sg}[x_i^+])$.
Output: trained one-step generator G_θ	

I Estimator and Implementation Formulas

For target references $\{x_j^+\}_{j=1}^{B_p}$ and model references $\{x_j^-\}_{j=1}^{B_q}$, define

$$\widehat{s}_p(y) = \frac{1}{B_p}\sum_{j=1}^{B_p}\varphi(x_j^+, y), \tag{110}$$

$$\widehat{s}_q(y) = \frac{1}{B_q}\sum_{j=1}^{B_q}\varphi(x_j^-, y), \tag{111}$$

$$\widehat{r}(y) = \frac{\widehat{s}_q(y)}{\widehat{s}_p(y)}. \tag{112}$$

For queries $Y_\ell \sim \varphi(x, \cdot)$, the direct general-$f$ estimator is

$$\widehat{\mathbf{u}}_f(x) = -\frac{1}{L}\sum_{\ell=1}^{L} f'(\widehat{r}(Y_\ell))\nabla_x \log\varphi(x, Y_\ell). \tag{113}$$

The corresponding specializations are:

$$\text{Gaussian general } f\text{:} \quad -\frac{1}{L}\sum_{\ell} f'(\widehat{r}(x+z_\ell))\frac{z_\ell}{h^2}, \tag{114}$$

$$\text{Gaussian KL:} \quad \frac{1}{L}\sum_{\ell}\left[\nabla\log\widehat{s}_p(x+z_\ell) - \nabla\log\widehat{s}_q(x+z_\ell)\right], \tag{115}$$

$$\text{vMF general } f\text{:} \quad -\frac{\kappa}{L}\sum_{\ell} f'(\widehat{r}(Y_\ell))\left(Y_\ell - (x^\top Y_\ell)x\right). \tag{116}$$

For KDE-L^2, use the induced kernel k_φ and the pairwise gradient estimator in Eq. (40).

I.1 Reference-sample bias of ratios and responses

Fix a query y and kernel parameters, and put $a = s_q(y)$, $b = s_p(y) > 0$, $r = a/b$. Write $\widehat{s}_q = a + \delta_q$ and $\widehat{s}_p = b + \delta_p$, with centered reference errors. Let $n_p, n_q \asymp n$, $V_q = \mathbb{E}\delta_q^2$, $V_p = \mathbb{E}\delta_p^2$, and $C = \mathbb{E}[\delta_q\delta_p]$. Assume these second moments are $O(n^{-1})$, the denominator along the Taylor segment is bounded below by a positive constant, and the expected third-order remainder is $o(n^{-1})$. Bounded kernel values, a uniformly positive kernel on a compact state space, and independent identically distributed reference samples give a sufficient setting: $\mathbb{E}\|\delta\|^3 = O(n^{-3/2})$. For response expansions below, additionally assume $\psi = f'$ has bounded third derivatives on the relevant ratio interval. Noncompact kernels require corresponding inverse-denominator and moment control.

A second-order expansion of $g(a, b) = a/b$ gives

$$\widehat{r} - r = \frac{\delta_q - r\delta_p}{b} + \frac{r\delta_p^2 - \delta_q\delta_p}{b^2} + R_n, \qquad \mathbb{E}|R_n| = o(n^{-1}). \tag{117}$$

Taking expectations yields the explicit bias

$$\mathbb{E}\widehat{r} - r = \frac{rV_p - C}{b^2} + o(n^{-1}). \tag{118}$$

For independent real and model reference batches, $C = 0$ and $V_p = \mathrm{Var}_{X\sim p}[\varphi(X, y)]/n_p$, proving Eq. (16). Except when this coefficient vanishes, the leading term is of order n^{-1}: the correct general statement is $O(n^{-1})$ bias with an $o(n^{-1})$ remainder. Unbiased component KDEs alone do not remove the ratio bias.

To propagate this result through the divergence response, set $g(a, b) = \psi(a/b)$. Its second derivatives are

$$g_{aa} = \frac{\psi''(r)}{b^2}, \quad g_{ab} = -\frac{\psi'(r) + r\psi''(r)}{b^2}, \quad g_{bb} = \frac{2r\psi'(r) + r^2\psi''(r)}{b^2}. \tag{119}$$

Consequently,

$$\begin{aligned}\mathbb{E}\psi(\widehat{r}) - \psi(r) = \frac{1}{2b^2}\big[&\psi''(r)V_q - 2\{\psi'(r) + r\psi''(r)\}C \\ &+ \{2r\psi'(r) + r^2\psi''(r)\}V_p\big] + o(n^{-1}).\end{aligned} \tag{120}$$

Thus even a linear first-order ratio approximation leaves a second-order response bias. The formula also covers correlated reference estimators through C; reference reuse should be analyzed with the corresponding covariance. When a clean particle belongs to its model reference batch, leave-one-out estimation removes its deterministic self-contribution.

I.2 Query averaging, field bias, and numerical evaluation

With exact densities, independent kernel queries give an unbiased Monte Carlo estimator of the clean-particle field. With empirical densities and queries independent of the reference batches, conditional expectation over queries instead gives

$$\mathbb{E}[\widehat{\mathbf{u}}_f(x) \mid \text{references}] = -\int_M \psi(\widehat{r}(y))\nabla_x\varphi(x, y)\,\mathrm{dvol}_g(y). \tag{121}$$

If the response expansion and its remainder admit integrable domination after multiplication by $\|\nabla_x\varphi(x, y)\|$, Eq. (120) can be integrated to obtain $\mathbb{E}\widehat{\mathbf{u}}_f(x) - \mathbf{u}_f(x) = O(n^{-1})$. Increasing the number L of independent queries reduces conditional Monte Carlo variance by L^{-1}, whereas the reference-batch bias is controlled by n_p, n_q. These are distinct approximation parameters.

For fixed bandwidth or concentration, the constants in the expansion depend on the kernel, the denominator, and the response derivatives. A bandwidth that changes with n requires tracking these constants. In numerical evaluation, log-sum-exp computes KDE ratios stably without changing the estimator. A fixed denominator floor or response clipping modifies the estimator and must be included in its bias analysis; the formulas above apply directly to the unclipped empirical ratio under the stated conditions.

Appendix K.1 specifies the target-derived kernel scales and corresponding query laws used in the experiments, including the distinction between fixed population scales and scales computed from the reference batch.

J Extended Experimental Protocols and Comparisons

We organize the experiments around four capabilities of the construction: changing the divergence response, adapting to geometry and data type, scaling one-step image generation, and matching distributions across evaluation representations. The cross-method comparisons use two fixed objectives: GFD-KL uses $\mathrm{KL}(s_q\|s_p)$, while GFD+mix uses $\frac{1}{2}\,\mathrm{KL}(s_p\|s_q) + \frac{1}{2}\chi^2_{1/2}(s_q\|s_p)$, with $\chi^2_{1/2}$ defined in Appendix D. The dedicated five-seed Swiss-roll study also tests KDE-L^2, additional individual divergences, and a quadratic–KL mixture (Appendix K.2).

Which experiments use an encoder? Our Euclidean toy, spherical toy, Earth-event, and promoter-DNA experiments are encoder-free: the GFD training field is constructed directly in the corresponding data space, without a pretrained or jointly learned feature extractor. DNA extends this evaluation beyond low-dimensional examples by optimizing directly in the high-dimensional space of length-1024 nucleotide sequences. CIFAR-10 and ImageNet-256 instead use feature encoders to construct the training signal. CIFAR-10 images are upsampled to 256×256 before feature extraction; both

tasks use multi-depth ViT CLS and patch features, and ImageNet additionally uses ConvNeXt feature maps at three resolutions (Appendix K.3). ImageNet uses frozen DINOv3 encoders. Here, encoder use refers to the training objective, not the neural generator itself or networks used only for evaluation.

J.1 Divergence responses under controlled comparisons

Euclidean space. Kernels and particle updates act directly on the two-dimensional target coordinates, with no feature encoder. This provides a controlled test of divergence responses without learned representation geometry.

We rebuild this comparison from the authors' original Drifting toy demo (Deng et al., 2026). We use its checkerboard and Swiss-roll samplers, 32-dimensional noise, three-layer width-256 SiLU MLP, Adam optimizer, and raw-model evaluation. The runs use learning rate 2×10^{-4}, 5,000 updates, and batches of 2,048. KGD, Drifting, W-Flow, GFD-KL, and GFD+mix use the same network and target batches. W-Flow calls its released optimal-transport loss with the released default radii and Sinkhorn settings (Han et al., 2026a).

We report the smallest raw-model SW_2 value recorded over training (Table 2). GFD+mix improves on GFD-KL on both tasks, while the relative ordering against other methods depends on the target: Drifting has the lowest checkerboard discrepancy and GFD+mix the lowest Swiss-roll discrepancy. The comparison tests the effect of changing the energy response under a shared backbone rather than establishing a uniformly best update rule. Sample plots are given in Appendix K.

Controlled CIFAR-10. We next test whether changing the divergence response has a meaningful effect on image generation under a matched training budget. All methods use the same DriftDiT model, frozen DINOv3 ViT-B/16 feature extractor, image preprocessing, optimizer, batch size, schedule, and evaluation code, while retaining their respective released training algorithms. GFD uses normalized spherical features, while the baselines retain their original Euclidean objectives (Appendix K.3). The comparison controls the representation and preprocessing without replacing each baseline by a common GFD loss implementation. GFD+mix has the lowest FID at every reported checkpoint, whereas GFD-KL has the highest IS after 10k steps (Figure 1). These distinct metric preferences show that the energy choice changes the learned distribution; they do not imply that one response dominates every criterion.

J.2 Geometry and data-type applicability

Spherical toy data. We first test the explicit vMF/spherical instance of Section 3.3, keeping particle updates intrinsic to the sphere. The kernels act on native spherical coordinates without a feature encoder.

We use the released KGD tasks, geometry, backbone, and metrics on $\mathbb{S}^2$. We compare intrinsic KGD with the two intrinsic vMF GFD objectives. All three runs use the same width-256 depth-5 ReLU MLP, Adam, learning rate 2×10^{-4}, 5,000 updates, and batches of 2,048. We again report the smallest raw-model SW_2 value, written simply as SW_2. KGD is best on spherical checkerboard, and GFD+mix is best on spherical Swiss roll. Sample plots are given in Appendix K.

Earth events. To test this geometric instance on real directional data, we use the four public Earth-event datasets from the KGD benchmark: Volcano, Earthquake, Fire, and Flood (Esteban-Casadevall et al., 2026; Mathieu & Nickel, 2020). GFD constructs its field directly on the spherical event coordinates, without an encoder or learned latent representation. All methods follow the benchmark split, backbone, budget, and evaluation code. The metric is the benchmark kernel discrepancy.

The preferred response varies across the four targets: GFD-KL has the lowest reported Volcano discrepancy, while GFD+mix has the lowest Flood discrepancy. This variation supports treating the divergence as a modeling choice rather than assuming a universal ordering.

Promoter DNA on the simplex. This experiment tests whether GFD can train directly in a high-dimensional data space without a feature encoder. The generator is optimized through the square-root categorical representation: the distributional energy and particle targets are constructed on the ambient product sphere from nucleotide coordinates, with no pretrained sequence encoder, learned feature

Table 10: Scaled CIFAR-10 FID comparison across training regimes.

Method	Training regime	NFE	FID ↓
Multi-step references			
EDM (Karras et al., 2022)	Score diffusion, direct	35	1.97
PFGM++ (Xu et al., 2023)	Physics-inspired diffusion, direct	35	1.74
One-step direct training without adversarial optimization			
iCT-deep (Song & Dhariwal, 2024)	Consistency training, direct	1	2.51
MeanFlow (Geng et al., 2025)	Average-flow training, direct	1	2.92
GFD+mix (ours)	KDE f-divergence flow, direct	1	<u>2.04</u>
One-step methods with generative pretraining, distillation, or adversarial optimization			
SiD (Zhou et al., 2024)	Score distillation	1	1.92
StyleGAN-XL (Sauer et al., 2022)	Adversarial training	1	1.85
CTM (Kim et al., 2024)	Trajectory distillation + DSM + adversarial	1	1.73
CCM (Liu et al., 2025)	Curriculum distillation + adversarial	1	1.64
GDD-I (Zheng & Yang, 2025)	Diffusion initialization + adversarial post-training	1	1.54
SiDA (Zhou et al., 2025)	Adversarial score distillation	1	1.396
Uni-Instruct (Wang et al., 2025)	Diffusion f-divergence instruction	1	**1.38**

extractor, or latent compression between the generated sequences and the GFD objective. Thus, the high-dimensional distribution-matching signal is not supplied by an external representation model.

To probe applicability beyond images and directional observations, we evaluate generative modeling on promoter DNA sequence design with discrete nucleotide tokens (A, C, G, T) for sequences of length $L = 1024$. Performance is evaluated using the k-mer distribution correlation (6-mer Corr. ↑) between real and generated sequence distributions. We report the inference computational cost in terms of Number of Function Evaluations (NFE).

GFD+mix reaches a 6-mer correlation of 0.96 with one function evaluation (Table 5). Together with the encoder-free toy and Earth-event results, this provides evidence that useful drifting-style distribution updates do not intrinsically require a learned feature encoder, including in a high-dimensional sequence task.

J.3 Scaling direct one-step image generation

Scaled CIFAR-10. Beyond the controlled comparison, we assess practical image-generation quality by scaling GFD+mix under the standard CIFAR-10 protocol while retaining one-step inference. Table 10 separates direct training from approaches using generative pretraining, distillation, or adversarial optimization. GFD+mix attains FID 2.04 at one NFE; the different training regimes provide context rather than a matched-budget comparison.

ImageNet-256 scope and model selection. We evaluate class-conditional one-step generation on ImageNet at 256×256 resolution. Our primary comparison is restricted to *directly trained* one-step generators: Drifting-L (Deng et al., 2026) is the closest pixel-space baseline to GFD-L, whereas W-Flow-B and W-Flow-L (Han et al., 2026a) provide approximately parameter-matched comparisons at the B and L scales, respectively, while operating in latent space.

We additionally evaluate AMFD-U (Liu et al., 2026) initialized from pMF-L (Lu et al., 2026) as a strong system-level reference. AMFD-U is not a direct-training baseline: it starts from a pretrained generator and performs conditional multi-representation post-training. We therefore separate direct training and post-training throughout the analysis.

Model size and training budget. Table 11 distinguishes trainable generator parameters from frozen representation parameters used for training. GFD uses DINOv3 (Siméoni et al., 2025) ViT-B/16 and ConvNeXt-B (173.2M frozen parameters in total), whereas Drifting-L and W-Flow use a 2.69B-parameter representation model. AMFD-U uses SigLIP2, MAE ViT-L, and Inception-v3 for post-training (755.1M frozen parameters). GFD-B/L are trained for 384 epochs (120k steps), compared with 640 epochs (100k steps) for Drifting-L and 1280 for W-Flow-B/L; AMFD-U uses 100 additional

post-training epochs. These settings combine one-step inference with fewer frozen representation parameters and fewer training epochs than the listed Drifting and W-Flow configurations.

Table 11: ImageNet-256 configurations and Inception metrics. Generator counts include the decoder where listed. Frozen counts describe training representations. Baseline FID values come from the original papers; GFD values come from the GFD runs. IS is independently recomputed. The ratio panel is a distinct checkpoint evaluation (Appendix L.2).

Method	Space	Generator (+decoder)	Frozen	Epochs (steps)
W-Flow-B	latent	133M+49M	2.69B	1280 (200k)
W-Flow-L	latent	463M+49M	2.69B	1280 (200k)
Drifting-L	pixel	463M	2.69B	640 (100k)
pMF-L/16	pixel	410.3M	90.2M	320 (400k)
GFD-B	pixel	133.9M	173.2M	384 (120k)
GFD-L	pixel	464.5M	173.2M	384 (120k)
AMFD-U (pMF-L)	pixel	410.3M	755.1M	+100 (125k)

Method	Training regime	NFE	FID ↓	IS ↑
W-Flow-B	direct	1	1.52	284.44
W-Flow-L	direct	1	1.35	294.56
Drifting-L	direct	1	1.61	346.10
pMF-L/16	direct	1	2.52	299.43
GFD-B	direct	1	3.14	293.18
GFD-L	direct	1	2.34	315.69
AMFD-U (pMF-L)	post-training	1	0.86	343.58

Generation quality at scale. Scaling GFD from B to L improves ImageNet FID from 3.14 to 2.34 and IS from 293.18 to 315.69 while keeping one-NFE inference. GFD-L does not have the lowest Inception FID among the directly trained references. Its IS lies above W-Flow-L and below Drifting-L, motivating a closer examination of what these generated distributions match in other evaluation representations.

J.4 Distribution matching across evaluation representations

We ask whether distribution matching learned in GFD's DINOv3 training space transfers to other representations. FID's Inception representation and Gaussian-moment approximation can miss perceptually relevant differences, as discussed by the CMMD and DINO-based FD studies (Jayasumana et al., 2024; Stein et al., 2023). We therefore retain FID and IS while also evaluating six encoder families not used to train GFD: SigLIP2, ConvNeXtV2-FCMAE, Inception-v3, DreamSim, EUPE-B, and Pixio-B. The longer metric rationale and complete readout protocol are in Appendix L; held-out status is method-specific, particularly for AMFD-U.

Metrics, normalization, and aggregation. Because raw discrepancy distances exhibit representation-dependent scales, direct cross-feature aggregation is ill-posed. We address this with the *validation-normalized ratio* $r_{\phi,D}$ in Eq. (18) for feature readout ϕ and discrepancy D,

where P_G, P_T, P_V denote the generated, training, and validation distributions, respectively. Here, $r_{\phi,D} = 1.0$ serves as the empirical reference baseline between independent data splits rather than a theoretical lower bound.

We use five discrepancies D: Fréchet distance (FD), sliced W_2^2 distance (SWD), RBF-MMD, and both directional KL estimates. The KL values are sliced density-estimation diagnostics rather than exact high-dimensional divergences. Appendix L.3 specifies the sample budgets and estimators.

Hierarchical aggregation. To prevent multi-readout encoders (e.g., CLS vs. AVG pooling) from over-dominating the benchmark, we use the two-stage aggregation in Eq. (18) across our six encoder families $\mathcal{F}$, with $R_{\text{avg5}} = \frac{1}{5}\sum_D R_D$.

Here, $R_D(G)$ provides a balanced family-level vote for each discrepancy D, while $R_{\mathrm{avg5}}(G)$ serves as a protocol-specific summary, not a representation-independent quality ordering. Appendix L.4 reports all 12 readouts for all five discrepancies.

What changes across evaluation representations? Table 7 shows that GFD-L has the lowest values for all five family-balanced discrepancies among the directly trained models evaluated here; GFD-B also has lower aggregate discrepancies than W-Flow-B. This differs from the standard Inception FID ordering in Table 11. The contrast is precisely why we report multiple representations and discrepancies: the panel detects distributional differences that a single Inception summary does not capture. It supports the effectiveness of these GFD instances under the stated protocol, not universal dominance across metrics or training regimes.

Post-training reference and representation dependence. AMFD-U has a lower aggregate discrepancy ($R_{\mathrm{avg5}} = 1.359$), but starts from pMF-L and uses multi-representation post-training. Its panel is not fully held out because its S+I+M training set comprises SigLIP2, Inception, and MAE, with the first two overlapping this evaluation panel. We therefore retain it as a separate system-level reference. Appendix L also examines post-training GFD with the AMFD representation set: FID improves from 2.34 to 1.98 before returning to 2.27, showing initial improvement followed by partial regression, consistent with a representation-shift interpretation.

K Experimental Details and Additional Results

The cross-method experiments use standard KL and an equal mixture of reverse KL and half-scaled Pearson χ^2; the five-seed Swiss-roll study additionally compares individual and mixed energies as detailed below. The response formulas are given in Appendices D and G. ImageNet-256 protocols and ratio results appear later in this appendix.

The two-dimensional experiment starts from the original Drifting demo. It uses the demo's checkerboard and Swiss-roll samplers, 32-dimensional Gaussian noise, a three-hidden-layer width-256 SiLU MLP, Adam, and raw-model evaluation. All ten runs use learning rate 2×10^{-4}, 5,000 updates, data and model batches of 2,048, and 5,000 evaluation particles. SW_2 uses 128 fixed random projections and is recorded every 500 updates.

Drifting uses the demo temperature 0.05. KGD uses its released Laplace-gradient toy settings: temperature 0.05. W-Flow uses its loss with temperature 0.05, 10 Sinkhorn iterations, non-squared distance cost, and an independent generated reference batch. GFD uses one exact Gaussian query per clean particle and scale, with current particles as leave-one-out model references. GFD+mix assigns equal weight to reverse KL and the half-scaled Pearson-χ^2 energy.

The spherical toy experiment uses the released KGD $\mathbb{S}^2$ tasks, manifold operations, width-256 depth-5 ReLU backbone, and metrics. All runs use 5,000 updates, batch size 2,048, Adam with learning rate 2×10^{-4}, gradient clipping at 1, and raw-model evaluation. SW_2 uses 4,096 samples and 128 random projections. KGD uses temperature 0.05. GFD uses target-scaled vMF kernels with dimensionless temperature 0.05 and exact Wood queries. We compare only intrinsic KGD and intrinsic GFD on these tasks.

The Earth runs use the benchmark 80/10/10 split, batch size 1,024, 30 epochs, and three evaluation repeats. These repeats concern evaluation, not independent training seeds.

Encoder-free training spaces. The Euclidean toy runs construct their GFD fields directly in two-dimensional data coordinates; the spherical toy and Earth-event runs use the native sphere. No learned feature encoder defines the training geometry in these experiments. The promoter-DNA experiment uses the deterministic square-root embedding of length-1024 categorical sequences into a product of spheres, without an external sequence encoder or a compressed latent space. This is a high-dimensional encoder-free test, complementing the low-dimensional geometric examples. The DINOv3-based nearest-neighbor analysis below and the ImageNet evaluation panel should not be confused with the spaces used to train these encoder-free instances.

K.1 GFD ESTIMATOR AND PARTICLE-TARGET CONFIGURATION

Table 12 collects the GFD-specific settings for the cross-method toy comparisons, Earth events, DNA, and image generation. The separate five-seed energy study is reported in Appendix K.2. Each component is indexed by its representation, spatial or sequence grouping scale, and kernel temperature; this index determines a smoothing kernel, a query law, and particle targets for the corresponding groups.

Table 12: GFD estimator settings. n_p, n_q count target and model references; L counts queries per clean particle and kernel component. $\mathcal{T} = \{0.03, 0.05, 0.3\}$. Euclidean targets are defined in Eq. (128); spherical targets use Eq. (129).

Experiment	Kernel	τ	(n_p, n_q)	L
Euclidean toy	Gaussian	0.05	$(2048, 2048)$	1
Spherical toy	vMF	0.05	$(2048, 2048)$	1
Earth events	vMF	0.05	$(1024, 1024)$	1
Promoter DNA	vMF / product-vMF	$\mathcal{T}$	$(64, 64)$	2
Controlled CIFAR-10	vMF / product-vMF	$\mathcal{T}$	$(64, 64)$	2
Scaled CIFAR-10	vMF / product-vMF	$\mathcal{T}$	$(64, 64)$	2
ImageNet-256	vMF / product-vMF	$\mathcal{T}$	$(64, 64)$	2

Target-adaptive, model-independent kernel scaling. Let $D_g(x, y)$ denote the dissimilarity for component g. We use squared Euclidean distance for Gaussian components and cosine dissimilarity for normalized spherical components:

$$D_g(x, y) = \begin{cases} \|x - y\|_2^2, & \text{Euclidean,} \\ 1 - x^\top y, & \text{spherical.} \end{cases} \tag{122}$$

For a group $\mathcal{G}$, first form $D_{\mathcal{G}}(X, Y) = \sum_{r \in \mathcal{G}} D_r(x_r, y_r)$. With target references X_i^+, estimate the scale from off-diagonal target pairs and define the kernel logit by

$$\widehat{\sigma}_{p,g} = \mathrm{Std}_{i \neq j}\left[D_g(X_i^+, X_j^+)\right], \qquad a_g = \tau \widehat{\sigma}_{p,g}, \qquad K_{\tau,g}(x, y) = e^{-D_g(x,y)/a_g}. \tag{123}$$

The normalized smoothing density is $\varphi_{\tau,g}(x, y) = K_{\tau,g}(x, y)/Z_g(x)$, where $Z_g(x)$ integrates the kernel over the query space. For the Euclidean and full-sphere kernels used here, Z_g is independent of x. Grouped components use one scale estimated from their summed dissimilarity; the factors are not separately standardized. This calibrates the variability of each joint logit using the target distribution. The formulas below take $\widehat{\sigma}_{p,g} > 0$.

Query distributions induced by the scaling. For a Euclidean component of dimension d, $\int_{\mathbb{R}^d} e^{-\|y-x\|^2/a_g}\, \mathrm{d}y = (\pi a_g)^{d/2}$. Comparing the exponent with a Gaussian density therefore gives

$$\varphi_{\tau,g}(x, \cdot) = \mathcal{N}\Big(x, \frac{a_g}{2} I\Big), \qquad Y = x + \sqrt{a_g/2}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{124}$$

The corresponding clean-particle kernel score is $2(Y - x)/a_g$. On $\mathbb{S}^{d-1}$, $e^{-(1-x^\top y)/a_g} = e^{-1/a_g} e^{x^\top y/a_g}$, so normalization gives

$$\varphi_{\tau,g}(x, y) = C_d(\kappa_g) e^{\kappa_g x^\top y}, \qquad \kappa_g = a_g^{-1}, \qquad Y \sim \mathrm{vMF}(x, \kappa_g). \tag{125}$$

Its intrinsic score is $\kappa_g\{Y - (x^\top Y)x\}$. For a spherical group with m factors, the summed cosine dissimilarity yields

$$\varphi_{\tau,g}(X, Y) = \prod_{r=1}^{m} C_{d_r}(\kappa_g) e^{\kappa_g x_r^\top y_r}, \qquad Y_r \mid X \sim \mathrm{vMF}(x_r, \kappa_g) \text{ independently.} \tag{126}$$

Thus the same group scale determines both the joint KDE and the concentration of its constituent queries. The Gaussian and Wood samplers in Appendix E apply with these parameters. Shared normalizing constants cancel in density ratios; the KDE-L^2 energy instead uses the normalized density and its induced kernel as defined in Appendix C.

Relation to the population construction. The population scale $\sigma_{p,g} = \mathrm{Std}_{X,X'\sim p}[D_g(X, X')]$ depends on the fixed target, so choosing it leaves the variation with respect to q unchanged. In the sample implementation, the target-derived scale is held fixed while constructing the field and its stopped target. If the same target batch selects the scale and supplies the KDE references, these random quantities are coupled; the fixed-kernel reference-bias calculation in Appendix I.1 describes the reference-error contribution for a prescribed kernel, rather than the entire random-scale estimator.

Reference sampling and leave-one-out estimation. The model references are generated samples and the target references are real samples. When the clean particle x_i belongs to the model-reference batch, its kernel contribution is excluded:

$$\widehat{s}_q^{(-i)}(y) = \frac{1}{n_q - 1} \sum_{j \neq i} \varphi_{\tau,g}(x_j, y), \qquad \widehat{s}_p(y) = \frac{1}{n_p} \sum_j \varphi_{\tau,g}(X_j^+, y). \tag{127}$$

Only the model batch contains this clean-particle self-term. Each query is sampled from the normalized kernel centered at x_i, with independent query random draws conditional on the selected kernel parameters and clean particle. DNA and image components use two queries, while the cross-method toy and Earth experiments use one. Product-kernel queries are conditionally independent across factors, whereas all coordinates of a reference group come from the same data sample and retain their joint dependencies.

Euclidean field normalization and finite step. For a batch of B particles, define $\mathrm{RMS}(u) = [B^{-1} \sum_i \|u(x_i)\|_2^2]^{1/2}$. Each active Euclidean field is normalized separately, and the normalized fields are averaged before multiplication by 0.1:

$$\widetilde{u}_j = \frac{\widehat{u}_j}{\mathrm{RMS}(\widehat{u}_j)}, \qquad u_{\mathrm{comp}} = \frac{1}{J} \sum_{j=1}^{J} \widetilde{u}_j, \qquad x_i^+ = x_i + 0.1\, u_{\mathrm{comp}}(x_i). \tag{128}$$

The displayed normalization applies to nonzero active fields; a zero field contributes zero. The composite field is not normalized again. Consequently, $\mathrm{RMS}(x^+ - x) = 0.1\,\mathrm{RMS}(u_{\mathrm{comp}}) \leq 0.1$, with its actual value determined by the alignment between components. Component normalization sets their effective step scales; the exact linear energy identity of Proposition 3.6 applies to the underlying fields before this numerical normalization.

Relation to prior drifting implementations. Component-wise scale balancing follows the normalization design of existing drifting systems. Drifting normalizes each feature/temperature drift before aggregation (Deng et al., 2026, Appendix A.6, Eqs. 23–26). The released W-Flow code likewise uses Drifting-style feature scaling and RMS-normalizes each regularization-scale velocity before summing (Han et al., 2026b). To compare with these baselines fairly, GFD follows this component-balancing convention, with its batch-vector RMS and equal averaging defined explicitly in Eq. (128). The prior methods use coordinate-mean RMS conventions, so this is a shared design principle rather than an assertion of identical normalization factors or aggregation weights. Balancing prevents raw field magnitude alone from determining a component's effective contribution. Proposition 3.6 describes the population fields before this numerical step.

Spherical finite-step targets. For a tangent field $u(x) \in T_x\mathbb{S}^{d-1}$, retain its direction and clip its geodesic displacement to $2^\circ = \pi/90$ radians. The single-sphere target is

$$\theta = \|u(x)\|_2, \qquad \bar{\theta} = \min(\theta, \pi/90), \qquad x^+ = \cos(\bar{\theta})x + \sin(\bar{\theta})\frac{u(x)}{\theta}, \tag{129}$$

with $x^+ = x$ when $\theta = 0$. Product-sphere coordinates use the corresponding sphere exponential maps. This constructs a unit-norm target while preserving the tangent direction. Equation (129) uses the closed-form exponential map with a clipped geodesic step; the generic first-order retraction in Appendix H is not the approximation used for these spherical targets. Distinct representation, grouping, and temperature components construct separate stopped targets and regression losses.

Energy composition and component-loss aggregation. Within a fixed spherical representation, group, and temperature, the GFD+mix objective and its field are

$$\mathcal{G}_{g,\mathrm{mix}} = \tfrac{1}{2}\,\mathrm{KL}(s_{p,g}\|s_{q,g}) + \tfrac{1}{2}\chi^2_{1/2}(s_{q,g}\|s_{p,g}), \tag{130}$$

$$u_{g,\mathrm{mix}} = \tfrac{1}{2}u_{g,\mathrm{RKL}} + \tfrac{1}{2}u_{g,\chi^2_{1/2}}. \tag{131}$$

Here $\chi^2_{1/2}$ includes the half-scaling in its definition, and $u_{g,\chi^2_{1/2}}$ is the field of that energy. This mixed field produces one geometric target for component g. For represented generator outputs $z_{\theta,g}$, the independent component losses are then summed with equal weights:

$$\mathcal{L}(\theta) = \sum_g \mathcal{L}_g(\theta), \qquad \mathcal{L}_g(\theta) = \mathbb{E}\Big[d^2_{M_g}\big(z_{\theta,g}, \mathrm{sg}[z_g^+]\big)\Big]. \tag{132}$$

This aggregation includes representation depths, feature maps, spatial or sequence group scales, and temperatures; ImageNet also includes both encoder families. In particular, the three temperatures specify three kernel components with separate targets, rather than one averaged smoothing kernel. The heterogeneous hierarchy supplies separate energy-derived stopped targets whose losses are aggregated in the neural objective; it is not represented as one population energy on a common state space.

K.2 Five-seed divergence and energy-composition study

This study compares forward KL, reverse KL, Pearson χ^2, KDE-L^2, reverse KL + χ^2, and KDE-L^2 + forward KL on Euclidean Swiss roll. Each row summarizes five training seeds. Table 3 reports mean and standard deviation of SW_2, together with the best and worst seed-level values. These are the dedicated five-seed energy comparison, separate from the cross-method results in Table 2.

Reference, estimator, and particle-target settings. All six energy choices use $n_p = n_q = 2048$ target and generated references. Each f-divergence field uses one Gaussian query per clean particle ($L = 1$). When a clean particle is part of the generated reference batch, its self-contribution is excluded using leave-one-out estimation as in Eq. (127). The KDE-L^2 component is computed directly from pairwise gradients of the KDE-induced kernel, using the estimator in Appendix C. In KDE-L^2 + forward KL, this pairwise component is combined with the one-query forward-KL field. Each active Euclidean field is RMS-normalized separately; the normalized fields are averaged with equal weights and multiplied by 0.1 to form the displacement in Eq. (128). The composite field is not normalized again.

Additional diagnostics. Table 13 reports the floor-normalized SW_2 diagnostic and W_2 evaluated with $n = 2048$, both as mean $\pm$ standard deviation across seeds. The reported MMD and MMD + forward KL rows use the KDE-induced kernel of Appendix C; they are labeled KDE-L^2 and KDE-L^2 + forward KL throughout the paper.

Table 13: Additional diagnostics for the five-seed Euclidean Swiss-roll study. SW_2 statistics and seed-level extrema are reported in Table 3; all discrepancy columns here are lower-is-better.

Energy	SW_2/floor	W_2 ($n = 2048$)
Forward KL	1.42 ± 0.53	$\mathbf{0.1084} \pm 0.0079$
Reverse KL	1.53 ± 0.50	0.1102 ± 0.0112
Pearson χ^2	$\mathbf{1.30} \pm 0.18$	0.1109 ± 0.0053
KDE-L^2	1.53 ± 0.43	0.1132 ± 0.0076
Reverse KL + χ^2	1.43 ± 0.41	0.1101 ± 0.0091
KDE-L^2 + forward KL	1.45 ± 0.49	0.1156 ± 0.0114

Effect of energy composition. Forward KL yields SW_2 0.0100 ± 0.0025, while KDE-L^2 gives 0.0107 ± 0.0016. Their combination attains 0.0099 ± 0.0008 and a worst-seed value of 0.0107, compared with 0.0130 and 0.0123 for the individual components. Pearson χ^2 reaches the same reported mean SW_2 and the best individual seed value of 0.0063, with a larger standard deviation

of 0.0043. Thus the combined quadratic–KL energy achieves a favorable balance of mean quality and seed stability. The additional diagnostics distinguish other properties: Pearson χ^2 has the lowest mean floor-normalized SW_2, while forward KL has the lowest mean W_2. These results illustrate the modeling value of choosing and combining energy responses within the same construction.

K.3 Multi-level image and sequence representations

Marginals and grouped dependencies. Single-component marginal distributions need not determine a joint distribution. We therefore match groups whose coordinates come jointly from the same real or generated sample. For m normalized components $Z = (z_1, \dots, z_m)$, the product-vMF kernel is

$$\varphi_{\mathrm{joint}}(Z, Y) = \prod_{r=1}^{m} \varphi_{\kappa_g}(z_r, y_r), \qquad Y_r \mid Z \sim \mathrm{vMF}(z_r, \kappa_g) \text{ independently.} \tag{133}$$

Conditional independence of the perturbations preserves the joint data law in the reference groups. Grouping therefore makes dependence structure accessible to the distributional objective. The single-component case is included once as the marginal component, and longer groups supply joint components. Representation depth, group size, and temperature index the separate losses in Eq. (132).

ViT features for CIFAR-10 and ImageNet. We use frozen DINOv3 ViT-B/16 features from transformer layers $3, 6, 9, 12$. At each depth, both the CLS token and all patch tokens participate in training. CIFAR-10 inputs are first upsampled to 256×256; the ViT patch grid is 16×16 for both image datasets. Each token feature is L_2-normalized for GFD. Every layer-specific CLS token forms a global single-sphere component. Its patch grid is partitioned into contiguous non-overlapping square groups with side lengths $1, 2, 4, 8, 16$. A side length s groups s^2 tokens into one sample on a product of spheres, with the single-token case included only once. Each depth, CLS or patch grouping scale, and temperature in $\{0.03, 0.05, 0.3\}$ defines a separate target and loss.

Table 14: Training feature hierarchy. ConvNeXt entries use the final output of each listed stage. Patch-group sizes are side lengths, not counts of regions. CIFAR-10 uses the ViT hierarchy; ImageNet uses both hierarchies. All features are normalized for the GFD spherical objectives.

Encoder	Depth / feature map	Grid	Components / group sides
ViT-B/16	Layer 3	16×16	CLS; $1, 2, 4, 8, 16$
ViT-B/16	Layer 6	16×16	CLS; $1, 2, 4, 8, 16$
ViT-B/16	Layer 9	16×16	CLS; $1, 2, 4, 8, 16$
ViT-B/16	Layer 12	16×16	CLS; $1, 2, 4, 8, 16$
ConvNeXt-B	Third-last stage	32×32	$1, 2, 4, 8, 16, 32$
ConvNeXt-B	Penultimate stage	16×16	$1, 2, 4, 8, 16$
ConvNeXt-B	Last stage	8×8	$1, 2, 4, 8$

Additional ConvNeXt hierarchy on ImageNet. ImageNet-256 also uses the final outputs of the last three spatial stages of frozen DINOv3 ConvNeXt-B, ordered from the third-last to the last stage, with feature-map resolutions 32×32, 16×16, and 8×8, respectively. At each resolution, we use non-overlapping groups at all power-of-two side lengths up to the full map size (Table 14). Normalized spatial features define spherical components, and grouped features use the product-vMF construction. Their stopped-target losses are added with the same component weights as the ViT losses. ImageNet training thus spans feature depth, spatial resolution, dependency scale, and kernel temperature.

Controlled CIFAR-10 geometry. GFD, Drifting, and W-Flow share the frozen DINOv3 ViT-B/16 feature extractor, image preprocessing, generator, optimizer, batch size, training schedule, and evaluation implementation. Each method retains its released training algorithm. GFD normalizes the extracted features and defines its vMF objectives on the resulting hyperspheres; Drifting and W-Flow use their original Euclidean objectives on the shared extracted representations. The controlled comparison therefore evaluates each training design with the same feature extractor and preprocessing.

Scale normalization and released baseline settings. The controlled CIFAR-10 comparison retains the released Drifting and W-Flow scale settings without an additional DINOv3-specific tuning procedure. Drifting rescales each feature by its empirical pairwise-distance scale and sets $\widetilde{\tau}_j = \tau\sqrt{C_j}$, where C_j is feature dimension (Deng et al., 2026, Appendix A.6, Eqs. 21–22). This removes feature magnitude and dimensionality from its temperature parameterization; the paper fixes $\tau \in \{0.02, 0.05, 0.2\}$ across configurations. Its feature/temperature drifts are normalized before aggregation. W-Flow uses alternating Sinkhorn scaling toward prescribed source and target marginals (Han et al., 2026a, Algorithm 2). Its reported entropy-regularization ablation varies ε around the recommended value and finds its best tested result at 0.05 (Han et al., 2026a, Table 2(b)). These normalization designs and reported default-parameter robustness motivate retaining the released configurations under the shared representation and budget. They provide an experimental protocol rationale, without implying an optimal parameter for every encoder.

DNA square-root coordinates. Each relaxed categorical nucleotide vector $\pi = (\pi_A, \pi_C, \pi_G, \pi_T) \in \Delta_3$ is represented by

$$z = \sqrt{\pi} = (\sqrt{\pi_A}, \sqrt{\pi_C}, \sqrt{\pi_G}, \sqrt{\pi_T}) \in \mathbb{S}^3_+, \qquad \mathbb{S}^3_+ = \{z \in \mathbb{R}^4 : z_a \geq 0,\ \|z\|_2 = 1\}. \tag{134}$$

This is an injective deterministic coordinate map with inverse $\pi_a = z_a^2$ on the positive orthant. A length-1024 sequence lies in $(\mathbb{S}^3_+)^{1024}$ as a subset of the ambient product sphere $M = (\mathbb{S}^3)^{1024}$. We partition the sequence into contiguous, non-overlapping groups of lengths $1, 4, 16, 64, 256, 1024$. Each group scale and each temperature in $\{0.03, 0.05, 0.3\}$ has its own vMF or product-vMF target and regression loss. The full-length group directly compares the joint sequence representation. This construction uses nucleotide coordinates without a learned sequence encoder.

DNA query support and stopped targets. Clean reference sequences have nonnegative square-root coordinates, while vMF queries are sampled on the full constituent spheres and may have negative coordinates. No truncation, rejection of negative coordinates, or folding into the positive orthant is applied. Hence the smoothing density is normalized on the full, boundaryless product sphere M, with its usual vMF normalizer and score. The spherical finite-step target is likewise retained in M even when it has negative coordinates and is used directly in the stopped-target regression loss. There is no projection of that target into $\mathbb{S}^3_+$. The generator representation remains the square root of a valid categorical vector; the full-sphere target is an auxiliary regression target, not a decoded nucleotide distribution. This separates the support of represented sequences from the ambient space used for smoothing and target construction. Discrete sequence evaluation retains the benchmark decoding and evaluation procedure.

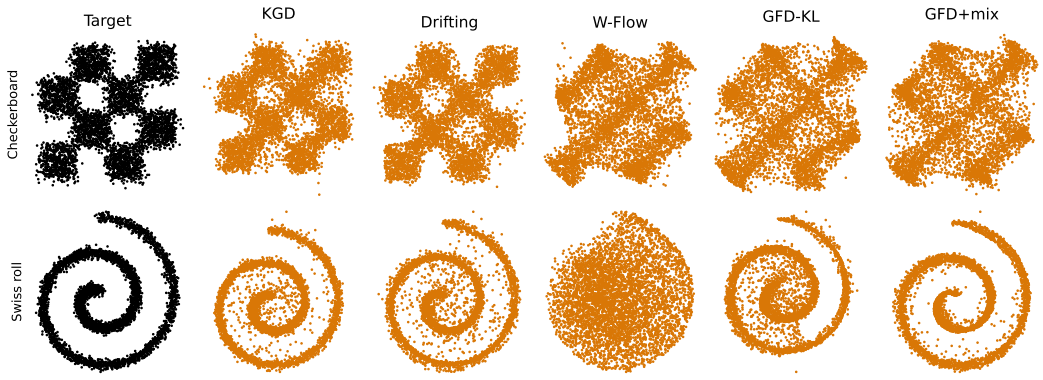


Figure 2: Euclidean toy samples. Rows show checkerboard and Swiss roll.

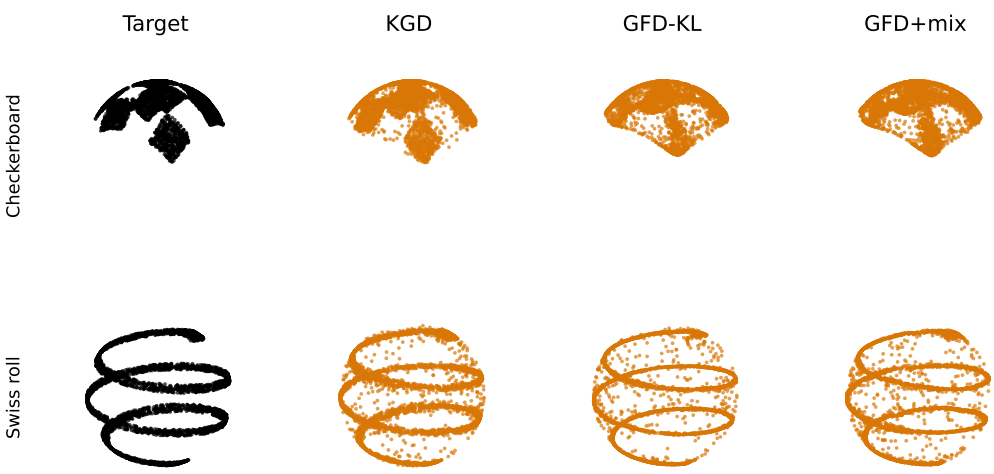


Figure 3: Spherical toy samples. Rows show checkerboard and Swiss roll.

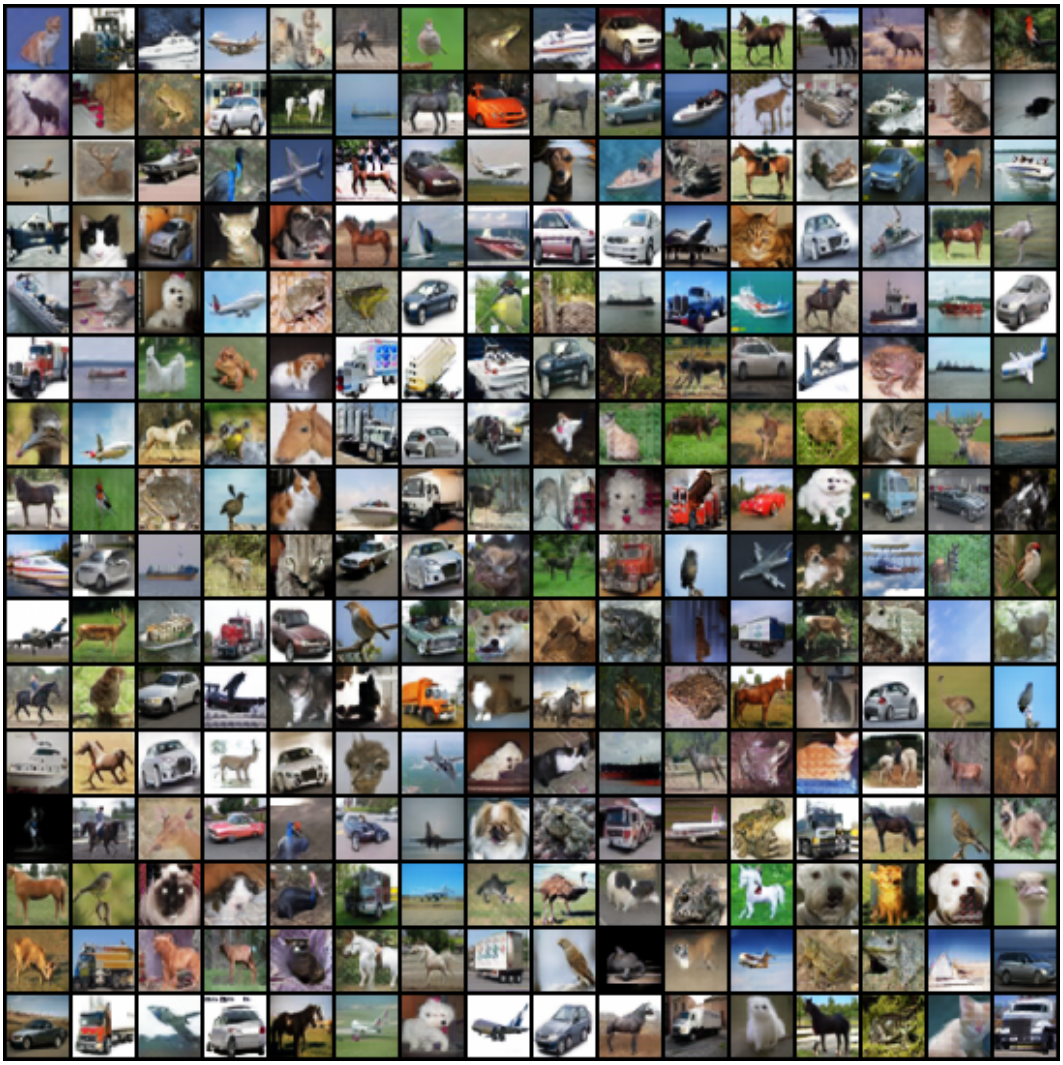


Figure 4: 256 samples generated from GFD+mix.

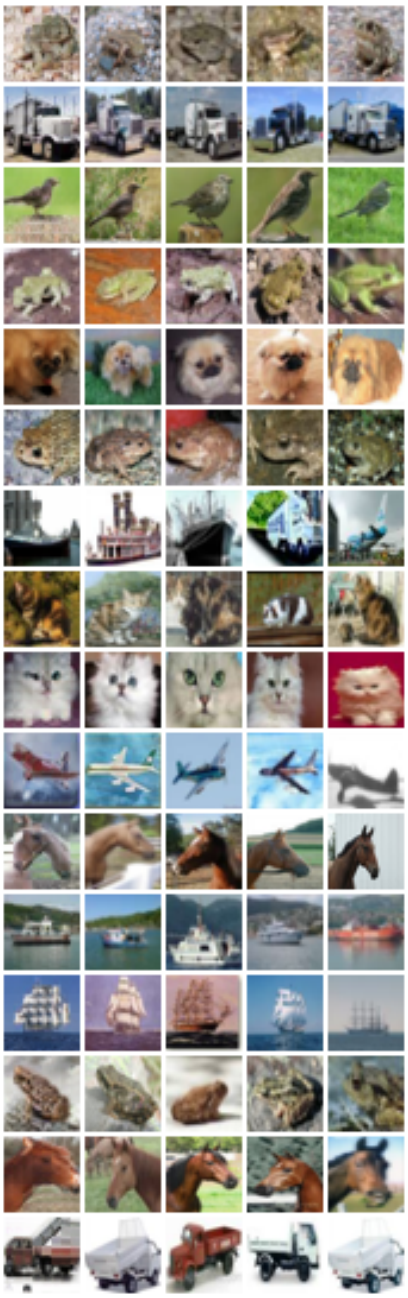


Figure 5: **Memorization analysis via nearest-neighbor retrieval.** For 16 randomly generated images (left column, one per row), we retrieve the top-4 most similar images from the CIFAR-10 training set (columns 2–5) using cosine similarity of DINOv3 ViT-B/16 CLS tokens.

K.4 ADDITIONAL IMAGENET-256 SAMPLES



Figure 6: GFD-L. Label: cardoon.



Figure 7: GFD-L. Label: dog.

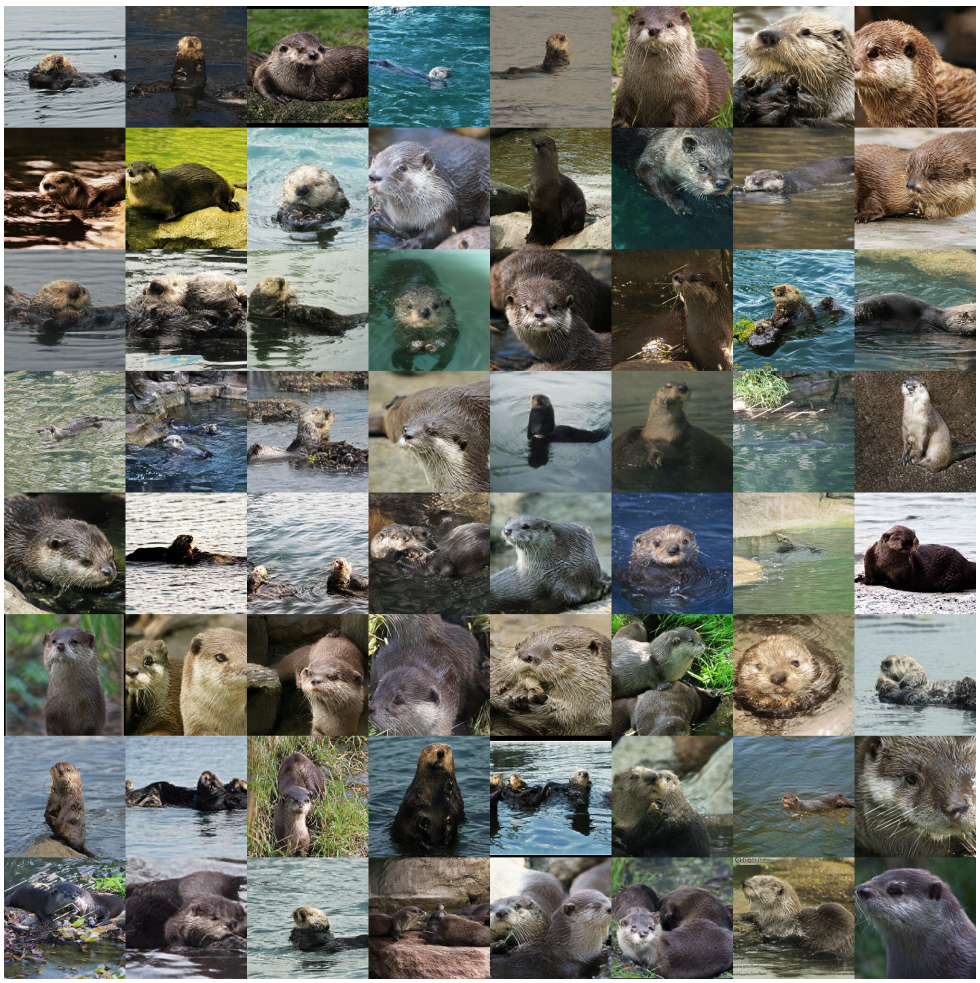


Figure 8: GFD-L. Label: otter.



Figure 9: GFD-L. Label: penguin.

L IMAGENET-256 EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

This section retains the settings needed to reproduce the multi-representation evaluation and reports only validation-normalized ratios; no raw feature-space distances are included.

L.1 RATIONALE FOR MULTI-REPRESENTATION EVALUATION

Why a multi-metric comparison? We retain FID and IS in Table 11 for continuity with standard ImageNet reporting, but we do not treat FID as a complete or representation-neutral ordering of generative models. Optimization in a representation space can also influence metrics computed in that space. FD-Loss (Yang et al., 2026) explicitly optimizes representation Fréchet distances and finds that the best-FID model need not be the visually best model; different training representations expose different properties of the generator. This motivates measuring both agreement and variation across evaluation spaces. The CMMD study of Jayasumana et al. (2024) identifies several relevant failure modes: FID is tied to the comparatively narrow semantics of supervised Inception-v3 features, approximates feature distributions using only Gaussian first and second moments, has poor sample complexity, and can fail to track controlled distortions, iterative quality improvements, or human preferences. Their CLIP-based MMD alternative illustrates that both the representation and the distributional discrepancy materially affect the resulting ranking. Complementarily, Stein et al. (2023) show that Inception-based metrics can undervalue the perceptual realism of diffusion models and that Fréchet distance computed with self-supervised DINOv2 features yields substantially richer evaluation and better agreement with human realism judgments; they also find that the information emphasized by a representation depends strongly on its architecture and pretraining objective. These observations imply that no single encoder–distance pair should be treated as an unquestioned oracle. We therefore retain conventional FID as one reference while grounding our comparative conclusions in agreement across heterogeneous held-out representation families and five complementary discrepancies.

Why these representation spaces? Evaluating a feature-trained generator only in its training representation risks circular evaluation bias. We therefore distinguish GFD's training feature space from a panel of other representation families; this measures distribution matching beyond the optimized representation. During training, GFD employs a dual-backbone DINOv3 setup (Siméoni et al., 2025) (ViT-B/16 and ConvNeXt-B), capturing rich global and dense self-supervised features while avoiding reliance on a single inductive bias.

To test transfer beyond the training representation, our evaluation panel incorporates six heterogeneous encoders spanning contrasting pretraining objectives: SigLIP2 (Tschannen et al., 2025) (vision–language alignment), ConvNeXtV2-FCMAE (Woo et al., 2023) (masked reconstruction), Inception-v3 (Szegedy et al., 2016) (supervised classification), DreamSim (Fu et al., 2023) (human-perceptual alignment), EUPE-B (Zhu et al., 2026) (multi-expert distillation), and Pixio-B (Yang et al., 2025) (pixel-supervised pretraining). This multi-objective benchmark avoids treating any single model as an unquestioned oracle, instead examining whether GFD's distributional matching transfers across disparate feature geometries and loss formulations.

L.2 METRIC PROVENANCE AND INDEPENDENTLY RECOMPUTED EVALUATION

The standard FID comparisons reproduce the baseline values reported in the original method papers; the GFD rows report the unchanged results of the GFD runs. In contrast, the multi-representation panel and its per-readout tables are independently computed from samples generated with publicly released baseline checkpoints and the GFD checkpoints, using a shared evaluation implementation. IS in the configuration table is also recomputed from the evaluated checkpoints and samples. Reported-paper FID and recomputed Inception-FD ratios are therefore distinct evaluations: the latter divide the independently measured Inception discrepancy by the validation-versus-training reference discrepancy in Eq. (18). They are not obtained by rescaling the published FID column. A denominator reconstructed by dividing a published FID by a recomputed ratio need not be common across these two tables. Within the independently recomputed panel, each readout–metric pair uses the same reference normalization across models. The panel protocol below specifies its own reference and sample budgets; we do not assume that every original paper used these same samples or reference statistics.

L.3 Representation and estimator protocol

The DINOv3 encoder families are used during GFD training; Table 15 lists the readouts used for evaluation. These DINOv3 evaluation readouts summarize the training encoder families and do not enumerate the multi-level features used by the GFD objective in Appendix K.3. The remaining six encoder families form the held-out evaluation panel for GFD and supply the family-balanced summary in Table 7. The complete breakdown contains 12 evaluation readouts.

Table 15: ImageNet-256 evaluation readouts: three summaries from the DINOv3 encoder families used in training and nine readouts from six held-out encoder families. This table specifies evaluation, while Table 14 specifies the multi-level training hierarchy.

Encoder	Encoder use in GFD	Evaluation readout(s)	Input
DINOv3 ViT-B/16 (Siméoni et al., 2025)	training	CLS, Patch-AVG	256
DINOv3 ConvNeXt-B (Siméoni et al., 2025)	training	Global Pool	224
SigLIP2 ViT-B/16-256 (Tschannen et al., 2025)	evaluation	CLS, Patch-AVG	256
ConvNeXtV2-FCMAE (Woo et al., 2023)	evaluation	Global Pool	224
Inception-v3 (Szegedy et al., 2016)	evaluation	Pool	299
DreamSim (Fu et al., 2023)	evaluation	Global	224
EUPE ViT-B (Zhu et al., 2026)	evaluation	CLS, Patch-AVG	224
Pixio ViT-B/16 (Yang et al., 2025)	evaluation	Global, Patch-AVG	256

All models are evaluated with the same feature-extraction and estimator implementations. FD uses all 50K generated samples against the 50K training-reference samples and matches feature means and covariances. SWD uses 5K samples and $L = 1000$ one-dimensional optimal-transport projections. RBF-MMD uses 5K samples with a median-heuristic bandwidth. Both directional KL estimates use one-dimensional sliced density estimation with 5K samples, $L = 1000$ projections, and 100 histogram bins. Each result is divided by the corresponding validation-versus-training reference distance according to Eq. (18).

L.4 Per-readout ratio results

Table 16 reports all 12×5 readout–metric combinations. Only ratios are shown. In the compact metric labels, FKL denotes $\mathrm{KL}(P_T \| P_G)$ and RKL denotes $\mathrm{KL}(P_G \| P_T)$; W-B and W-L denote W-Flow-B and W-Flow-L.

Table 16: Validation-normalized ratios $r_{\phi,D}$ for all 12 readouts and the five main-text metrics. Lower is better.

Representation	Metric	W-B	Drift-L	W-L	pMF-L	GFD-B	GFD-L	AMFD-U
DINOv3 ViT-B CLS	FD	11.999	6.688	10.569	7.309	3.072	2.184	2.905
DINOv3 ViT-B CLS	SWD	26.519	13.793	22.728	13.160	1.721	1.320	5.387
DINOv3 ViT-B CLS	MMD	39.671	17.949	32.957	17.051	1.803	1.336	5.877
DINOv3 ViT-B CLS	FKL	2.378	1.641	2.139	1.571	1.047	0.998	1.205
DINOv3 ViT-B CLS	RKL	1.424	1.158	1.340	1.166	0.995	0.981	1.046
DINOv3 ViT-B Patch-AVG	FD	15.400	8.435	13.404	10.279	2.617	2.030	3.408
DINOv3 ViT-B Patch-AVG	SWD	21.595	11.749	18.327	12.076	1.269	1.052	4.414
DINOv3 ViT-B Patch-AVG	MMD	25.943	15.238	22.171	15.103	1.148	0.912	5.274
DINOv3 ViT-B Patch-AVG	FKL	1.742	1.455	1.620	1.348	0.853	0.823	1.105
DINOv3 ViT-B Patch-AVG	RKL	1.828	1.393	1.732	1.498	1.245	1.237	1.170
SigLIP2-B/16-256 CLS	FD	7.483	11.428	6.483	7.267	5.033	3.913	2.873
SigLIP2-B/16-256 CLS	SWD	10.922	18.139	8.974	10.575	5.400	3.961	3.858
SigLIP2-B/16-256 CLS	MMD	14.947	24.679	12.061	13.691	7.039	4.921	4.708
SigLIP2-B/16-256 CLS	FKL	2.141	2.580	1.965	2.025	1.620	1.470	1.319
SigLIP2-B/16-256 CLS	RKL	1.313	1.715	1.213	1.325	1.062	0.991	1.048
SigLIP2-B/16-256 Patch-AVG	FD	12.709	11.258	10.835	9.326	5.119	3.710	2.978
SigLIP2-B/16-256 Patch-AVG	SWD	20.699	18.521	16.750	12.480	4.812	3.102	3.626
SigLIP2-B/16-256 Patch-AVG	MMD	26.742	25.032	21.789	15.419	5.977	3.611	4.321
SigLIP2-B/16-256 Patch-AVG	FKL	2.171	2.223	1.938	1.548	1.206	1.099	1.183

Representation	Metric	W-B	Drift-L	W-L	pMF-L	GFD-B	GFD-L	AMFD-U
SigLIP2-B/16-256 Patch-AVG	RKL	1.432	1.326	1.323	1.253	1.068	1.021	1.008
DINOv3 ConvNeXt-B Pool	FD	16.897	2.480	14.172	4.109	0.915	0.886	1.160
DINOv3 ConvNeXt-B Pool	SWD	16.943	4.665	13.206	4.377	0.654	0.730	1.789
DINOv3 ConvNeXt-B Pool	MMD	19.734	9.671	16.088	7.801	0.365	0.318	3.076
DINOv3 ConvNeXt-B Pool	FKL	2.853	1.660	2.487	1.596	0.848	0.791	1.170
DINOv3 ConvNeXt-B Pool	RKL	2.086	1.208	1.798	1.128	1.107	1.218	1.029
ConvNeXtV2-FCMAE Pool	FD	2.249	0.809	1.877	1.274	2.225	1.639	0.623
ConvNeXtV2-FCMAE Pool	SWD	2.196	0.625	1.544	0.949	1.214	0.989	0.568
ConvNeXtV2-FCMAE Pool	MMD	2.764	0.699	1.977	1.249	1.594	1.252	0.547
ConvNeXtV2-FCMAE Pool	FKL	1.072	0.888	0.985	0.953	0.976	0.929	0.899
ConvNeXtV2-FCMAE Pool	RKL	1.072	1.058	1.086	1.041	1.064	1.084	1.058
Inception-v3 Pool	FD	1.097	1.058	1.011	1.478	2.351	1.756	0.644
Inception-v3 Pool	SWD	0.835	1.046	0.842	1.861	1.476	1.589	0.566
Inception-v3 Pool	MMD	0.855	1.057	0.905	2.102	1.878	1.861	0.505
Inception-v3 Pool	FKL	1.069	1.112	1.091	1.228	1.110	1.150	0.935
Inception-v3 Pool	RKL	0.907	0.908	0.903	0.903	0.960	0.924	1.039
DreamSim Global	FD	3.189	2.700	2.889	3.944	4.711	3.356	1.378
DreamSim Global	SWD	4.562	3.000	3.812	6.438	4.312	3.250	1.375
DreamSim Global	MMD	5.009	3.457	4.230	6.825	4.974	3.485	1.311
DreamSim Global	FKL	1.178	1.130	1.139	1.250	1.194	1.134	1.000
DreamSim Global	RKL	1.077	1.020	1.057	1.149	1.068	1.033	1.019
EUPE-B CLS	FD	7.370	5.375	6.302	5.766	4.913	4.067	1.489
EUPE-B CLS	SWD	9.244	5.964	7.674	5.736	4.203	3.697	1.204
EUPE-B CLS	MMD	10.328	6.822	8.572	6.825	4.585	3.890	1.313
EUPE-B CLS	FKL	1.418	1.270	1.329	1.266	1.177	1.138	1.013
EUPE-B CLS	RKL	1.335	1.199	1.271	1.184	1.126	1.095	1.000
EUPE-B Patch-AVG	FD	8.511	6.368	7.325	7.092	5.468	4.509	1.557
EUPE-B Patch-AVG	SWD	12.142	7.718	10.309	8.194	5.240	4.337	1.417
EUPE-B Patch-AVG	MMD	12.402	8.576	10.457	8.695	5.420	4.400	1.439
EUPE-B Patch-AVG	FKL	1.552	1.370	1.452	1.349	1.132	1.093	0.977
EUPE-B Patch-AVG	RKL	1.550	1.324	1.466	1.396	1.291	1.249	1.056
Pixio-B Global	FD	19.787	18.786	16.160	12.547	11.853	7.363	2.195
Pixio-B Global	SWD	21.942	20.365	17.077	10.346	9.058	5.588	1.479
Pixio-B Global	MMD	28.886	26.370	22.561	13.922	11.168	6.986	1.953
Pixio-B Global	FKL	2.577	2.401	2.292	1.866	1.604	1.383	1.018
Pixio-B Global	RKL	1.894	1.804	1.645	1.327	1.332	1.151	1.032
Pixio-B Patch-AVG	FD	26.870	23.724	22.085	16.057	15.284	9.517	1.836
Pixio-B Patch-AVG	SWD	31.576	28.020	25.658	15.266	11.934	7.339	1.742
Pixio-B Patch-AVG	MMD	34.807	32.169	28.188	17.058	12.786	8.001	1.868
Pixio-B Patch-AVG	FKL	2.730	2.715	2.422	1.910	1.633	1.358	1.075
Pixio-B Patch-AVG	RKL	2.271	2.050	2.016	1.595	1.522	1.307	0.995

L.5 Sensitivity to removal of an evaluation family

We assess whether the family-balanced aggregate depends on a single representation by omitting each of the six evaluation families in turn and applying the same within-family readout averaging and five-metric averaging to the remaining five families. The three evaluation readouts from DINOv3 training encoder families remain excluded in every case. This is a recombination of the reported per-readout ratios in Table 16, using their displayed three-decimal values; it requires no new samples or model training. The resulting aggregates are therefore approximate at the displayed precision, rather than recomputations from unrounded raw distances.

GFD-L remains the lowest-discrepancy directly trained model in all six omissions, including removal of Pixio, whose ratios have comparatively large magnitudes. Thus its aggregate lead is not contingent on any single evaluation family. This robustness concerns the specified arithmetic aggregation only: equal family weights do not remove all scale effects, and neither this test nor the full aggregate establishes dominance in every representation. In particular, Drifting-L has lower individual FD ratios

Table 17: Leave-one-family-out R_{avg5} (lower is better), computed from the reported rounded ratios. Bold identifies the best directly trained model; AMFD-U is a post-training reference.

Omitted family	W-B	Drift-L	W-L	pMF-L	GFD-B	GFD-L	AMFD-U
SigLIP2	5.949	4.910	4.939	4.094	3.499	**2.607**	1.092
ConvNeXtV2	7.586	7.085	6.307	5.373	3.983	**2.927**	1.482
Inception-v3	7.770	7.041	6.416	5.289	3.955	**2.871**	1.483
DreamSim	7.360	6.796	6.081	4.808	3.615	**2.672**	1.387
EUPE	6.643	6.329	5.483	4.642	3.574	**2.573**	1.381
Pixio	4.493	4.080	3.804	3.754	2.702	**2.163**	1.326

in ConvNeXtV2-FCMAE, Inception-v3, and DreamSim. AMFD-U remains a separate post-trained reference and has the lowest overall aggregate in each omission.

L.6 Post-training with the AMFD representation set

We further post-train the pretrained GFD-L generator with the same feature-extractor set used by AMFD-U—SigLIP2-So400m, MAE ViT-L, and Inception-v3—while retaining the GFD training objective. This continuation experiment changes the representation geometry from DINOv3 to the AMFD representation set while retaining the GFD objective. As shown in Table 18, FID initially improves from 2.34 to a best observed value of 1.98, but later rises to 2.27 at the final checkpoint.

Table 18: FID trajectory when post-training pretrained GFD-L with the AMFD representation set. Lower is better.

Stage	FID ↓
Pretrained GFD-L	2.34
Best during post-training	**1.98**
Final post-training checkpoint	2.27

The early improvement indicates that the AMFD-aligned feature objectives can refine a strong GFD initialization over a limited post-training horizon. The non-monotone trajectory, however, suggests negative transfer under a persistent representation shift. GFD-L is pretrained with DINOv3 ViT-B/16 and ConvNeXt-B features, so its generator is specialized to the feature geometry that shaped the original clean-particle field. Replacing that supervisory geometry with the AMFD representation set changes the effective optimization target: continued updates increasingly favor the new feature space while eroding structure retained from the pretraining space. We interpret the later FID increase as consistent with representation-induced forgetting and objective shift; the trajectory alone does not isolate this mechanism. The final FID of 2.27 remains better than the initial 2.34, so the observed degradation is relative to the intermediate optimum, not a loss of every post-training gain. This result motivates early stopping, mixed-feature objectives, or an explicit anchor to the pretraining representations when transferring a feature-space-trained generator across supervisory geometries.

M Additional Related Work and Field Comparisons

One-step generation and discrepancy learning. GANs and VAEs learn direct generators using adversarial and variational objectives (Goodfellow et al., 2014; Kingma & Welling, 2014). Diffusion, score-based models, and flow matching construct inference-time dynamics, while consistency models amortize these dynamics into one or a few evaluations (Ho et al., 2020; Song et al., 2021; Lipman et al., 2023; Song et al., 2023). Generator Matching maps Markov-process operators to generative models (Holderrieth et al., 2025). Sinkhorn generator learning and MMD-GAN optimize sample discrepancies, with the latter also learning a representation adversarially (Genevay et al., 2018; Li et al., 2017). GFD combines explicit distributional energies with stopped-target training: geometry, smoothing, and energy determine a clean-particle field, and the corresponding algorithm transfers its updates to a one-step generator.

From Drifting to a directly estimable energy family. Drifting evolves generated particles through an attraction–repulsion field and regresses the targets into the generator (Deng et al., 2026); W-Flow uses Wasserstein/Sinkhorn transport (Han et al., 2026a). Gretton et al. (2026) distinguish a smoothed-density score field from the clean-measure variation of a smoothed KL energy. Their Section 4.1 also describes energy-derived drifted targets and stopped-target generator training. GFD develops a specific direct KDE route for the family $D_f(\mathcal{S}_k q \| \mathcal{S}_k p)$: scalar density-ratio responses, adjoint kernel-query estimation, explicit geometry-dependent fields, and nonnegative energy composition. This differs from obtaining general f-divergence extensions through dual/critic formulations. The quadratic KDE energy supplies a complementary ratio-free component with the induced-kernel identity proved in Appendix C.

Table 19: Operational comparison of closely related constructions. Entries describe the formulations and experiments emphasized in the cited works, not limits on possible extensions. All connect particle updates to generator training. The columns group population/training choices, response/estimator choices, and geometry/complementary energies.

Method	Population construction and training	General f route and field estimator	Geometry and other energies
Drifting (Deng et al., 2026)	Attraction–repulsion field; stopped drift targets	Affinity-weighted sample interactions; a general direct KDE-f response is not specified	Euclidean feature spaces; MMD drift derivation in Appendix C.2; multi-temperature aggregation
Gretton et al. (2026)	Smoothed-KL and Sinkhorn-proxy analysis; energy-derived stopped targets	Adjoint clean-measure KDE-KL; general f extensions via dual/critic formulations	Euclidean examples; MMD and sliced-Wasserstein flows; comparisons across energy choices
KGD (Esteban-Casadevall et al., 2026)	Kernel-score difference applied to generated particles	Score/KL-type response; direct kernel-gradient estimator at clean particles	Intrinsic manifold and categorical constructions; characteristic-kernel identification
W-Flow (Han et al., 2026a)	Energy-functional framework, instantiated with Sinkhorn divergence; stopped WGF targets	Main algorithm uses empirical OT maps; no general adjoint KDE-ratio estimator is specified	Euclidean OT; MMD/KL energy baselines; velocity-level guidance
GFD	Direct $D_f(\mathcal{S}_k q \| \mathcal{S}_k p)$ and quadratic KDE energy; stopped targets	General scalar KDE-ratio responses; adjoint kernel-query fields and induced-kernel pairwise estimator	Normalized manifold kernels; KDE-L^2; explicit and evaluated nonnegative response compositions

Table 19 locates GFD's operational contribution within the existing energy-to-generator approach. Clean-measure KDE-KL and classical MMD flow are established ingredients (Gretton et al., 2026; Arbel et al., 2019); GFD integrates them into a directly estimable KDE f-divergence family with compatible geometry, explicit estimators, and compositional training. In particular, Drifting already discusses MMD, and W-Flow already formulates a general energy-functional framework. The distinguishing construction here is the direct KDE-ratio/adjoint route for general responses, together with its quadratic KDE counterpart.

Clean-particle evolution versus smoothed-law evolution in KGD. KGD (Esteban-Casadevall et al., 2026) applies a KDE score difference to clean generated particles. The distinction can be seen directly in Euclidean space with translation-invariant smoothing. Write $v = \nabla \log s_p - \nabla \log s_q$. If clean particles evolve with v, then

$$\partial_t q = -\operatorname{div}(qv) \quad \Longrightarrow \quad \partial_t s_q = -\operatorname{div}\big[\mathcal{S}_k(qv)\big]. \tag{135}$$

Here $\mathcal{S}_k$ acts componentwise on the vector measure qv. In contrast, evolving the smoothed law directly with the same velocity gives

$$\partial_t s_q = -\operatorname{div}(s_q v). \tag{136}$$

In general $\mathcal{S}_k(qv) \neq s_q v$, so the two equations describe different dynamics. KGD's descent argument for the smoothed variable does not provide the clean-measure descent identity for Eq. (135). Substituting its velocity into the derivative of the clean-measure KDE-KL objective instead gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{KL}(\mathcal{S}_k q \| \mathcal{S}_k p) = \int \langle \nabla \mathcal{A}_k \log(s_q/s_p), v\rangle \,\mathrm{d}q, \tag{137}$$

which has no automatic negative-square form. This is the precise heuristic step in using a smoothed-law field as the clean-particle gradient of that energy. GFD uses $\mathbf{u}_{\mathrm{KL}} = -\nabla \mathcal{A}_k \log(s_q/s_p)$ and obtains $-\int \|\mathbf{u}_{\mathrm{KL}}\|^2 \,\mathrm{d}q$ from the same objective.

KGD also gives a characteristic-kernel argument for distributional identification when its field vanishes everywhere. That property concerns global zeros of the field; it does not equate the two evolution equations above. GFD establishes both the clean-measure variational identity and global-field identification under the conditions of Theorem 3.3, and extends the construction to non-KL responses and quadratic KDE energies. Its spherical construction follows by intrinsic differentiation of the normalized kernel, with exact directional query sampling specified in Appendix E.1.

Variational and kernel connections. Wasserstein flows of distributional energies and f-divergences have a broad literature (Jordan et al., 1998; Ambrosio et al., 2008; Liu et al., 2023; Wang & Yu, 2026; Caucheteux et al., 2026). Score-difference flow uses $\nabla \log p - \nabla \log q$ for KL (Weber, 2023); Gaussian KDE-KL yields its kernel-averaged clean-particle form. The Gaussian KDE score is also the mean-shift direction (Cheng, 1995; Comaniciu & Meer, 2002). The induced-kernel representation of KDE-L^2 recovers classical MMD dynamics (Arbel et al., 2019; Sriperumbudur et al., 2010; 2011). Concurrent analyses investigate score interpretations, spectral behavior, conservative fields, and finite-batch corrections (Lai et al., 2026; Turan et al., 2026; Franz et al., 2026; Zhang et al., 2026). These connections place the direct KDE construction within existing variational theory while specifying its trainable non-KL fields and common generator-training algorithm.