Coreset-Induced Conditional Velocity Flow Matching

Zihua She
Department of Statistics
Purdue University
West Lafayette, IN 47907
she8@purdue.edu

Jianxi Su
Terry College of Business
University of Georgia
Athens, GA 30602
jianxi.su@uga.edu

Xiao Wang
Department of Statistics
Purdue University
West Lafayette, IN 47907
wangxiao@purdue.edu

Abstract

We propose Coreset-Induced Conditional Velocity Flow Matching (CCVFM), a generative model that augments hierarchical rectified flow with a data-informed source distribution. Standard flow matching regresses only the conditional mean of the velocity; hierarchical flow matching instead models the full conditional velocity law with a second, inner flow in velocity space, but this inner flow is asked to transport isotropic Gaussian noise to a multimodal target velocity distribution from scratch. Our key observation is that this inner source can be replaced by a closed-form surrogate built from a coreset of the target. CCVFM first compresses the target into weighted atoms using an entropic Sinkhorn coreset and lifts them to a Gaussian mixture. The induced conditional velocity law is then a closed-form Gaussian mixture that can be sampled without a learned neural sampler. A lightweight correction flow, trained from this surrogate source, then refines the remaining surrogate-to-target residual rather than learning an entire noise-to-data map. We prove that the surrogate transport cost equals the target–surrogate Wasserstein gap under an explicit compression assumption, whereas the noise-source analogue has a dimension-scale lower bound. We further characterize the conditional second moment of the training target when the source velocity is drawn from the surrogate law independently of the target velocity, and show that the part of it that depends on the source is small when the surrogate conditional law is close to the true conditional velocity law in mean and covariance. Empirically, on MNIST, CIFAR-10, ImageNet-32, and CelebA-HQ, the proposed method reaches competitive few-step generation under matched architectures. Code is available at https://github.com/ziwaa-se/CCVFM.

1 Introduction

Flow matching [17, 18, 1] has emerged as a simple and competitive paradigm for generative modeling. Given a source $X_0 \sim \rho_0$ and a target $X_1 \sim \rho_1$, one interpolates $X_t = (1-t)X_0 + tX_1$, defines the velocity $V = X_1 - X_0$, and regresses a field $u_\theta(x,t)$ on V under a squared loss. The minimizer is the *conditional mean* $u^\star(x,t) = \mathbb{E}(V \mid X_t = x)$, which suffices to match the marginal evolution but *collapses the entire conditional law* $\pi(v \mid x,t)$ *to a single direction*. Whenever the same intermediate state x is consistent with several endpoint completions, mean-field learning blurs the local transport structure and forces long integration paths (high NFE) at inference to compensate. The NFE cost of

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

few-step generation therefore has a structural component that no amount of solver engineering can remove: the field itself has been averaged.

The closest prior attempt to break this bottleneck is the *hierarchical rectified flow* (HRF2) of Zhang et al. [31], which models the full conditional law $\pi(v \mid x, t)$ via a second-level flow in velocity space. HRF2 proves a key structural identity: when ρ_1 is a GMM, the induced conditional velocity law is itself a closed-form GMM [Thm. 1 and Cor. 1 of 31], and we leverage this identity extensively. Its limitation is the choice of source: the correction flow starts from an isotropic $v_0 \sim \mathcal{N}(0, I)$ and must learn the *entire* transport from noise to a potentially multimodal velocity law, so the correction step carries most of the load and low-NFE generation remains hard.

We introduce Coreset-Induced Conditional Velocity Flow Matching (CCVFM), which keeps HRF2's hierarchical structure but replaces its isotropic source with a data-informed surrogate in closed form. Stage I fits K weighted atoms using entropic Sinkhorn and lifts them to a GMM $\widetilde{\rho}_1 = \sum_k w_k \mathcal{N}(\mu_k, \Sigma_k)$. Stage II plugs $\widetilde{\rho}_1$ into the closed-form velocity identity, giving a zero-parameter mixture sampler for $\widetilde{\pi}(v \mid x, t)$. Stage III trains a correction flow whose source velocity is drawn through the mixture representation of $\widetilde{\pi}$ rather than from $\mathcal{N}(0, I)$, so that it learns a surrogate-to-target residual instead of the full noise-to-data map.

The theoretical comparison is sharp at the generation boundary $t = 0$, where $\pi(v \mid x_0, 0) = \rho_1(x_0 + v)$. The CCVFM correction cost equals $W_2(\rho_1, \widetilde{\rho}_1)$ and is $O(K^{-1/d})$ under our explicit Stage-I compression assumption; the corresponding noise-source cost has a positive dimension-scale lower bound. In addition, write $C = (X_t, t)$ and let $\mu_\pi(C), \Sigma_\pi(C)$ and $\mu_{\mathrm{sur}}(C), \Sigma_{\mathrm{sur}}(C)$ denote the conditional means and covariances of $\pi(\cdot \mid C)$ and $\widetilde{\pi}(\cdot \mid C)$. When the source velocity is drawn from $\widetilde{\pi}(\cdot \mid C)$ independently of the target velocity, the conditional second moment of the Stage III training residual equals $\operatorname{tr} \Sigma_\pi(C) + \operatorname{tr} \Sigma_{\mathrm{sur}}(C) + \|\mu_\pi(C) - \mu_{\mathrm{sur}}(C)\|^2$. Its source-dependent excess, i.e. the part beyond the intrinsic spread $\operatorname{tr} \Sigma_\pi(C)$ shared by every source, is small whenever the surrogate conditional law is close to the true conditional velocity law in mean and covariance, in contrast to the independent-Gaussian-source baseline whose source-dependent excess is $\|\mu_\pi(C)\|^2 + d$.

Relation to standard flow matching. With a single outer step (J=1, our default and the setting of our theory), CCVFM is equivalent at the population optimum to data-space flow matching from $\widetilde{\rho}_1$ to ρ_1 under the same endpoint coupling (§3.4). It shares with flow matching / conditional flow matching (FM/CFM) the straight interpolation, squared loss, and Euler sampling; its novelty lies in the coreset-GMM surrogate source and the resulting surrogate-to-data residual-correction task. Samplers with $J > 1$ outer steps are a secondary extension (Appendix §B).

Contributions. (1) We introduce a coreset-to-velocity pipeline that replaces the inner Gaussian source in hierarchical flow matching by a compact, analytically sampleable coreset-GMM surrogate, which provides a more informative source distribution. (2) We prove a transport-task reduction (Theorem 1): the CCVFM correction-source cost vanishes with K under the Stage-I compression assumption, while the HRF2 (noise-source) cost is bounded below by a positive constant in K and n. We further establish a second-moment identity for the surrogate source (Theorem 3): a source velocity drawn from $\widetilde{\pi}(\cdot \mid C)$ through its mixture weights, independently of the target velocity, has marginal exactly $\widetilde{\pi}(\cdot \mid C)$, and its conditional training-residual second moment admits a closed-form expression in the conditional means and covariances of $\pi(\cdot \mid C)$ and $\widetilde{\pi}(\cdot \mid C)$, whose source-dependent excess is small whenever the surrogate is close to the target in mean and covariance. We validate both identities of Theorem 3 by Monte Carlo on seven datasets (Appendix §D). (3) We obtain strong few-step image results: MNIST $\mathrm{FID}_{50k} = 0.75$, CIFAR-10 $\mathrm{FID}_{50k} = 6.35$, ImageNet-32 $\mathrm{FID}_{50k} = 8.76$, and CelebA-HQ 256 $\mathrm{FID}_{28k} = 4.17$, all at 51 NFE. (4) Building on existing nearest-neighbour-based generative-model diagnostics [2, 19, 3, 4], we adapt the 1-NN primitive into a 1-NN Inception-feature goodness-of-fit diagnostic showing no preferential closeness to training data.

2 Background

Rectified flow [18] and conditional flow matching [17] sample $X_0 \sim \mathcal{N}(0, I_d)$ and $X_1 \sim \rho_1$, set $X_t = (1-t)X_0 + tX_1$, and regress a field on $V = X_1 - X_0$. The optimum $\mathbb{E}[V \mid X_t = x]$ matches marginals but discards the full conditional law, which motivates velocity-distribution methods such as HRF2 [31]. HRF2 learns $\pi(v \mid x_t, t)$ through a second flow with Gaussian inner source and uses

the identity

$$\pi(v \mid x_t, t) \propto \rho_0\big(x_t - tv\big)\, \rho_1\big(x_t + (1-t)v\big), \tag{1}$$

whose GMM specialization gives explicit component means, covariances, and weights whenever ρ_1 is a GMM. With $V_1 = X_1 - X_0$, HRF2 draws $V_0 \sim \mathcal{N}(0, I_d)$ independently of V_1 and regresses an inner field $f_\theta(V_\tau, \tau, X_t, t)$ on $V_1 - V_0$, where $V_\tau = (1-\tau)V_0 + \tau V_1$. The remaining difficulty is that HRF2's inner flow starts from isotropic noise and must cross mixture modes.

Weighted coresets compress data into atoms $\{(w_k, \mu_k)\}_{k=1}^K$ that approximate ρ_1 in transport distance [7, 6]. Classical quantization [10] gives the $O(K^{-1/d})$ benchmark for globally optimal K-atomic approximants under standard assumptions. For the deployed entropic-Sinkhorn/GMM output we use the corresponding rate only through the explicit Stage-I compression assumption.

3 CCVFM

CCVFM has three stages (Figure 1): build a coreset-GMM surrogate $\widetilde{\rho}_1$, read off its closed-form conditional velocity law $\widetilde{\pi}(v \mid x_t, t)$, and train a correction flow whose source is built from $\widetilde{\pi}(\cdot \mid C)$, so the learned residual is a surrogate-to-target correction rather than a noise-to-data map.

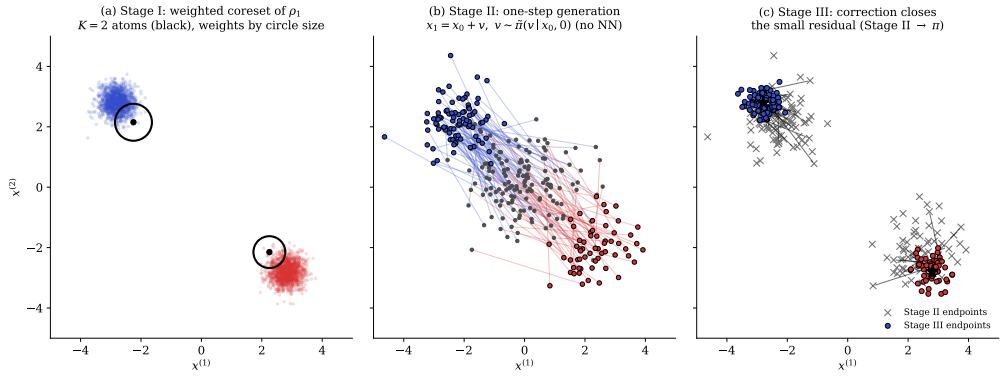


Figure 1: CCVFM pipeline: an entropic-Sinkhorn coreset is lifted to a GMM surrogate $\widetilde{\rho}_1$; the induced law $\widetilde{\pi}(v \mid x_t, t)$ is a closed-form GMM; a correction flow with a source built from $\widetilde{\pi}$ refines the surrogate-to-target residual.

3.1 Stage I: Sinkhorn coreset and GMM lift

Given samples $x_1, \ldots, x_n \sim \rho_1$ and budget $K \ll n$, Stage I approximates $\hat{\rho}_{1,n} = n^{-1}\sum_i \delta_{x_i}$ by $\sum_k w_k \delta_{\mu_k}$. The coupling $T \in \mathbb{R}_+^{n \times K}$ records soft assignments from data to anchors, with marginals $\sum_k T_{ik} = 1/n$ and $\sum_i T_{ik} = w_k$; T is used only inside Stage I to update the anchors and weights.

Concretely, we compute (w, μ, T) jointly by minimizing the Kullback-Leibler-regularized entropic objective

$$\min_{\{\mu_k, w_k\},\, T_{ik} \geq 0} \sum_{i,k} T_{ik} \|x_i - \mu_k\|^2 + \lambda\, \mathrm{KL}\big(T \,\big\|\, \tfrac{1}{n}\mathbf{1}_n \otimes w\big), \qquad \sum_k T_{ik} = \tfrac{1}{n},\ \ \sum_i T_{ik} = w_k, \tag{2}$$

with $\mathrm{KL}(T\|P) = \sum_{i,k} T_{ik} \log(T_{ik}/P_{ik})$ for the prior measure $P_{ik} = \frac{1}{n} w_k$. Holding (μ, w) fixed and optimizing in T subject to the row constraints yields the closed-form update $T_{ik} \propto w_k \exp(-\|x_i - \mu_k\|^2/\lambda)$ (the GMM E-step used in Algorithm 1), since the column marginal $\sum_i T_{ik} = w_k$ is generated rather than constrained. We alternate this T-update with barycentric anchor updates $\mu_k \leftarrow w_k^{-1} \sum_i T_{ik} x_i$, $w_k \leftarrow \sum_i T_{ik}$, at $O(nK)$ cost per outer iteration. The atom measure is then lifted to $\widetilde{\rho}_1 = \sum_k w_k \mathcal{N}(\mu_k, \Sigma_k)$, where $\Sigma_k = L_k L_k^\top + \sigma_k^2 I_d$ is fit from weighted residuals by SVD (component-specific noise variance, matching Algorithm 1). Smoothing makes $\tilde{\pi}$ well-defined; the deployed compression rate is handled explicitly in Assumption 1.

3.2 Stage II: closed-form conditional velocity law

Once $\widetilde{\rho}_1$ is available we can read off the induced conditional velocity law without any further learning. Plugging $\widetilde{\rho}_1$ into the identity (1) and completing the square in v yields:

$$\begin{aligned}\widetilde{\pi}(v \mid x_t, t) &= \sum_{b=1}^{K} \gamma_b(x_t, t)\, \mathcal{N}\big(v; m_b(x_t, t), \Lambda_b(t)\big), \\ A_b(t) &= t^2 I + (1-t)^2 \Sigma_b^{-1},\ \Lambda_b(t) = A_b(t)^{-1}, \\ b_b(x_t, t) &= t x_t - (1-t)\Sigma_b^{-1}(x_t - \mu_b),\ m_b(x_t, t) = \Lambda_b(t)\, b_b(x_t, t),\end{aligned} \tag{3}$$

with closed-form weights $\gamma_b(x_t, t) \propto w_b\, \mathcal{N}\big(x_t; t\mu_b, (1-t)^2 I + t^2 \Sigma_b\big)$. At the generation boundary $t = 0$, (3) simplifies drastically:

$$\widetilde{\pi}(v \mid x_0, 0) = \widetilde{\rho}_1(x_0 + v) = \textstyle\sum_b w_b\, \mathcal{N}(v; \mu_b - x_0, \Sigma_b). \tag{4}$$

Sampling from (4) at $t = 0$ requires only a categorical draw $b^* \sim \mathrm{Cat}(w_1, \ldots, w_K)$ followed by a Gaussian sample, with no learned function evaluation. The low-rank factor lets each draw run in $O(dr)$ via $v \leftarrow \mu_{b^*} - x_0 + L_{b^*} z + \sigma_{b^*} \xi$ with $z \in \mathbb{R}^r, \xi \in \mathbb{R}^d$ standard normal.

3.3 Stage III: correction flow with a surrogate source

Stage II samples from $\widetilde{\pi}$ track π only up to the surrogate gap $W_2(\rho_1, \widetilde{\rho}_1)$, which is still nonzero on realistic image data. Stage III closes the gap with a second flow-matching stage *in velocity space*: we train a correction network whose source is built from the closed-form surrogate law $\widetilde{\pi}(\cdot \mid C)$, which is the source law used at inference, and we run the two flows as a nested sampler at inference.

Given (x_0, x_1) with $X_t = (1-t)x_0 + t x_1$ and $V_1 = x_1 - x_0$, each training step draws a source velocity V_0 from the mixture representation of $\widetilde{\pi}(\cdot \mid X_t, t)$ (the two-step draw is given after (5)), picks $\tau \sim \mathcal{U}[0, 1]$ and forms $V_\tau = (1-\tau)V_0 + \tau V_1$, and regresses a correction net $f_\theta(v, \tau, x_t, t)$ toward $V_1 - V_0$:

$$\mathcal{L}(\theta) = \mathbb{E}\big\| f_\theta(V_\tau, \tau, X_t, t) - (V_1 - V_0)\big\|^2. \tag{5}$$

Equation (5) has the same form as HRF2's inner objective, but the source V_0 is built from the informative $\widetilde{\pi}(\cdot \mid C)$ rather than $\mathcal{N}(0, I)$, so the learnable residual is a surrogate-to-target correction rather than a noise-to-data transport.

The classical flow-matching choice for the source V_0 is the independent isotropic Gaussian, $V_0 \sim \mathcal{N}(0, I_d)$ drawn independently of $V_1 \sim \pi$, which yields a conditional residual $\mathbb{E}[\|V_1 - V_0\|^2 \mid X_t, t] = \mathrm{tr}\, \Sigma_\pi(X_t, t) + \|\mu_\pi(X_t, t)\|^2 + d$, lower-bounded by $c_\gamma D_\pi^2 + d$ whenever $\pi(\cdot \mid X_t, t)$ is multimodal with cross-mode diameter D_π (see Section A.5). This full cross-mode variance is what the correction network would have to absorb, and is exactly the transport task CCVFM was designed to avoid.

We instead draw V_0 through the mixture representation of Eq. (3). With $C = (X_t, t)$, a component label B is drawn first, followed by the Gaussian step $V_0 \mid B = b, C \sim \mathcal{N}\big(m_b(C), \Lambda_b(t)\big)$. When the label is drawn from the surrogate weights, independently of V_1, $B \mid C \sim \mathrm{Cat}\big(\gamma_1(C), \ldots, \gamma_K(C)\big)$, marginalizing over the component label gives

$$p(V_0 = v \mid C) = \sum_{b=1}^{K} \gamma_b(C) \mathcal{N}\big(v; m_b(C), \Lambda_b(t)\big) = \widetilde{\pi}(v \mid C).$$

We call this the *direct surrogate-source sampler*: its source marginal is exactly the surrogate law $\widetilde{\pi}(\cdot \mid C)$, the same source law used at inference, and it is the draw analyzed in Theorem 3; other label draws keep the same Gaussian step (Remark 20). Our image experiments anchor the label to the training image; the theoretical properties of related anchored couplings are analyzed in Wang et al. [29]. At the generation boundary $t = 0$, this reduces to

$$B \sim \mathrm{Cat}(w_1, \ldots, w_K), \qquad V_0 \mid B = b, X_0 = x_0 \sim \mathcal{N}(\mu_b - x_0, \Sigma_b),$$

so the source law at $t = 0$ is $\widetilde{\pi}(v \mid x_0, 0) = \widetilde{\rho}_1(x_0 + v)$ exactly. The resulting correction flow is trained from a data-adapted surrogate source rather than from an isotropic Gaussian source.

3.4 Generation procedure

Our theory (§4) analyzes the $J = 1$ case at $t = 0$, where $\pi(v \mid x_0, 0) = \rho_1(x_0 + v)$ collapses the conditional velocity law to a translated copy of the target. We therefore recommend $J = 1$ as the default, with the inner correction budget L the only knob; this estimator enjoys the closed-form guarantees of Theorem 1.

One-step recommended sampler ($J = 1$). Given the trained correction field f_θ and a budget $L \geq 1$: (1) draw $X_0 \sim \mathcal{N}(0, I_d)$; (2) sample $b^* \sim \mathrm{Cat}(w_1, \ldots, w_K)$ and $V \sim \mathcal{N}(\mu_{b^*} - X_0, \Sigma_{b^*})$, which realizes $V \sim \widetilde{\pi}(\cdot \mid X_0, 0)$ via (4); (3) for $\ell = 0, \ldots, L-1$: $V \leftarrow V + (1/L)\, f_\theta(V, \ell/L, X_0, 0)$; (4) return $\hat{X}_1 \leftarrow X_0 + V$. The Stage II step (2) is a categorical draw plus a Gaussian sample, with no learned-network evaluation; we count it as a single zero-parameter NFE for benchmarking equivalence with diffusion baselines that include the initial denoising step. Hence total $\mathrm{NFE} = L + 1$, used uniformly across all CCVFM FID rows in Section 5 and matching Tables 1 and 3.

$J = 1$ as data-space flow matching. With $Y_0 = X_0 + V_0$ and $Y_1 = X_1$, the $t = 0$ slice of (5) is flow matching in data space from the law of Y_0 to ρ_1 (with X_0 as an extra input); since (Y_0, Y_1) is independent of X_0, its population optimum coincides with that of data-space flow matching under the same endpoint coupling. The two views differ only at the estimator level, and Theorem 1 depends only on the source marginal (Appendix §B.5).

General $J > 1$ (optional). Because $\widetilde{\pi}(\cdot \mid x, t)$ is closed-form at every t, the same network also defines a nested sampler on an outer grid $0 = t_0 < \cdots < t_J = 1$ with total $\mathrm{NFE} = J(L + 1)$ (Appendix §B). This variant is *not* covered by the theory; for the networks trained over $t \in [0, 1]$ reported here, $J = 1$ matched or beat $J \geq 2$ on every dataset we tested (Table 8). With a network trained only on the slices $t \in \{0, \frac{1}{2}\}$ that $J = 2$ visits, $J = 2$ improves MNIST FID at equal learned-network evaluations (Table 5). Characterizing when $J > 1$ pays off is left to future work. Pseudocode for both variants is in Appendix §B.

4 Theory

Our theoretical analysis has two parts. §4.1 quantifies the conditional transport task that the correction net must solve at $t = 0$, contrasting the deployed Stage I source with the HRF2 noise source. §4.2 verifies the exact source marginal of the direct surrogate-source sampler of §3.3 and compares its conditional training-target scale with that of an independent Gaussian source. We analyze the $J = 1$ sampler of §3.4, where the velocity identity collapses and computations become closed-form.

Define the *conditional transport cost* of a correction stage with source $Q(\cdot \mid x, t)$ at $t = 0$ by

$$D(Q) \;:=\; \Big(\mathbb{E}_{X_0 \sim \rho_0}\big[\, W_2^2\big(Q(\cdot \mid X_0, 0), \pi(\cdot \mid X_0, 0)\big)\big]\Big)^{1/2}. \tag{6}$$

All proofs are deferred to Appendix §A.

Assumption 1 (Stage-I compression for the deployed surrogate)**.** Let μ_{in} denote the probability measure compressed by Stage I, and let $\widetilde{\rho}_1 = \sum_{k=1}^{K} w_k\, \mathcal{N}(\mu_k, \Sigma_k)$ be the K-component surrogate returned by the deployed entropic-Sinkhorn/GMM procedure. We assume that there exists a constant $C_Q < \infty$, independent of K, such that

$$W_2(\mu_{\mathrm{in}}, \widetilde{\rho}_1) \leq C_Q K^{-1/d}. \tag{7}$$

Here W_2 is the Euclidean 2-Wasserstein distance on $\mathbb{R}^d$, and we use the assumption only in the population surrogate-gap statement (Theorem 1(i)) and in the K-rate of Corollary 2, where $\mu_{\mathrm{in}} = \rho_1$. This is an assumption on the implemented Stage-I output. Classical quantization theory gives the corresponding $K^{-1/d}$ benchmark for globally optimal K-atomic quantizers on bounded d-dimensional domains, but it does not by itself prove (7) for finite-iteration entropic Sinkhorn followed by a GMM lift.

The nominal exponent d makes (7) pessimistic for images; for data near a lower-dimensional set we read it with an effective dimension $d_{\mathrm{eff}} \ll d$. A transport-cost proxy gives $d_{\mathrm{eff}} \approx 12$ for MNIST and ≈ 15 for CIFAR-10 (an empirical diagnostic, Appendix §D).

4.1 Bounding the surrogate gap

At $t = 0$, (1) simplifies to $\pi(v \mid x_0, 0) = \rho_1(x_0 + v)$: the conditional velocity law is the target shifted by $-x_0$. For any surrogate of the form $Q(v \mid x_0, 0) = q(x_0 + v)$, the Wasserstein distance to $\pi(\cdot \mid x_0, 0)$ equals $W_2(q, \rho_1)$ and is *independent of* x_0, so CCVFM and HRF2 can be compared by a single scalar per source.

Theorem 1 (Transport-task reduction)**.** *Let* $Q_{\mathrm{CCVFM}}(v \mid x_0, 0) = \widetilde{\rho}_1(x_0 + v)$ *and* $Q_{\mathrm{HRF2}}(v \mid x_0, 0) = \mathcal{N}(v; 0, I_d)$*. Then*

(i) $D(Q_{\mathrm{CCVFM}}) = W_2(\rho_1, \widetilde{\rho}_1)$*. Under Assumption 1 with* $\mu_{\mathrm{in}} = \rho_1$*,* $D(Q_{\mathrm{CCVFM}}) \le C_Q K^{-1/d}$*.*

(ii) $D(Q_{\mathrm{HRF2}}) \ge \sqrt{d} \cdot \big(\sqrt{\sigma_1^2 + 1} - 1\big)$*, where* $\sigma_1^2 = \mathbb{E}\|X_1\|^2/d$ *is the per-coordinate second moment of* ρ_1*. This lower bound is strictly positive whenever* $\sigma_1 \neq 0$ *and* constant in K and n*.*

Consequently, whenever the coreset error tends to zero with K*,* $D(Q_{\mathrm{CCVFM}})/D(Q_{\mathrm{HRF2}}) \to 0$ *as* $K \to \infty$ *for every fixed nondegenerate target: CCVFM's correction net is asked to solve a transport task that is arbitrarily smaller than HRF2's.*

Theorem 1 certifies that, asymptotically, CCVFM's correction net is asked to solve a transport task that vanishes with K, while HRF2's correction-source cost remains bounded below by a positive constant in K and n.

Corollary 2 (NFE sufficient scaling)**.** *Let* $D := D(Q)$*, and let* M *bound the* L^2 *norm of the correction field along the exact trajectory, and* L_v *and* L_τ *be its Lipschitz constants in* v *and in* τ*. Under the space-time regularity assumptions in Appendix §A.4 and the residual flattening conditions* $M = O(D)$*,* $L_v = O(D)$*, and* $L_\tau = O(D^2)$ *for the correction ODE, a sufficient number of Euler function evaluations to reduce the* L^2 *discretization error below* $\varepsilon_{\mathrm{disc}}$ *is*

$$\mathrm{NFE}_Q(\varepsilon_{\mathrm{disc}}) = O\left(1 + \frac{D(Q)^2}{\varepsilon_{\mathrm{disc}}}\right).$$

Consequently, for CCVFM, with $\Delta_K := D(Q_{\mathrm{CCVFM}}) = W_2(\rho_1, \widetilde{\rho}_1)$*,*

$$\mathrm{NFE}_{\mathrm{CCVFM}}(\varepsilon_{\mathrm{disc}}) = O\left(1 + \frac{\Delta_K^2}{\varepsilon_{\mathrm{disc}}}\right).$$

If, in addition, Assumption 1 gives $\Delta_K = O(K^{-1/d})$*, then* $\mathrm{NFE}_{\mathrm{CCVFM}}(\varepsilon_{\mathrm{disc}}) = O\Big(1 + \frac{K^{-2/d}}{\varepsilon_{\mathrm{disc}}}\Big)$*. For HRF2, if* d *and* ρ_1 *are fixed and the corresponding regularity constants are* $O(1)$*, then* $\mathrm{NFE}_{\mathrm{HRF2}}(\varepsilon_{\mathrm{disc}}) = O\Big(1 + \frac{1}{\varepsilon_{\mathrm{disc}}}\Big)$*.*

Corollary 2 is a conditional sufficient result: the flattening conditions are assumptions on the learned field, not consequences of a small D (Remark 12). It is consistent with CCVFM's empirical behavior: matched architectures but far fewer correction steps to reach the same FID when the learned residual field is flatter for the data-informed source. Without the flattening condition, the generic sufficient bound is $O(1 + (L_v D + L_\tau)/\varepsilon_{\mathrm{disc}})$ with the regularity constants treated as problem-dependent quantities.

4.2 Exact surrogate-source sampling and residual scale

Stage III trains the correction net f_θ by regressing on the residual $V_1 - V_0$ at random interpolation time τ (Eq. (5)). The conditional second moment $\mathbb{E}[\|V_1 - V_0\|^2 \mid C]$ with $C = (X_t, t)$ controls the scale of the flow-matching regression target. Under the direct surrogate-source sampler of §3.3, V_0 is drawn from the surrogate law $\widetilde{\pi}(\cdot \mid C)$ itself, independently of V_1.

Theorem 3 (Direct surrogate-source second moment vs. independent Gaussian source)**.** *Fix* $C = (X_t, t)$ *and let* $V_1 \sim \pi(\cdot \mid C)$*. Define* $\mu_\pi(C) := \mathbb{E}[V_1 \mid C], \Sigma_\pi(C) := \mathrm{Cov}(V_1 \mid C), S_\pi(C) := \mathrm{tr}\,\Sigma_\pi(C)$*. Let the surrogate law be* $\widetilde{\pi}(v \mid C) = \sum_{b=1}^K \gamma_b(C)\mathcal{N}\big(v; m_b(C), \Lambda_b(t)\big)$*. Define its conditional mean by* $\mu_{\mathrm{sur}}(C) := \sum_{b=1}^K \gamma_b(C) m_b(C)$*, and decompose its conditional covariance trace as* $\mathrm{tr}\,\Sigma_{\mathrm{sur}}(C) = A_{\mathrm{sur}}(C) + B_{\mathrm{sur}}(C)$*, where* $A_{\mathrm{sur}}(C) := \sum_{b=1}^K \gamma_b(C)\,\mathrm{tr}\,\Lambda_b(t)$ *is the average within-component covariance scale, and* $B_{\mathrm{sur}}(C) := \sum_{b=1}^K \gamma_b(C)\|m_b(C) - \mu_{\mathrm{sur}}(C)\|^2$

is the weighted between-component spread of the surrogate source. Equivalently, $B_{\mathrm{sur}}(C) = \frac{1}{2}\sum_{b=1}^{K}\sum_{b'=1}^{K}\gamma_b(C)\gamma_{b'}(C)\|m_b(C) - m_{b'}(C)\|^2$. *Finally, let* $M_{\mathrm{sur}}(C) := \|\mu_\pi(C) - \mu_{\mathrm{sur}}(C)\|^2$.

(i) Direct surrogate-source sampler. *If*

$$B \mid C \sim \mathrm{Cat}\big(\gamma_1(C), \ldots, \gamma_K(C)\big), \qquad V_0 \mid B = b, C \sim \mathcal{N}\big(m_b(C), \Lambda_b(t)\big),$$

and this draw is conditionally independent of V_1 *given* C, *then* $V_0 \mid C \sim \widetilde{\pi}(\cdot \mid C)$, *and*

$$\mathbb{E}\big[\|V_1 - V_0\|^2 \,\big|\, C\big]_{\mathrm{sur}} = S_\pi(C) + A_{\mathrm{sur}}(C) + B_{\mathrm{sur}}(C) + M_{\mathrm{sur}}(C). \tag{8}$$

(ii) Independent Gaussian source. *If* $V_0 \sim \mathcal{N}(0, I_d)$ *independently of* V_1 *given* C, *then*

$$\mathbb{E}\big[\|V_1 - V_0\|^2 \,\big|\, C\big]_{\mathrm{Gauss}} = S_\pi(C) + \|\mu_\pi(C)\|^2 + d. \tag{9}$$

Moreover, suppose $\pi(\cdot \mid C)$ *admits a latent mixture representation with component label* Z, *weights* $\gamma_b^\pi(C) := \Pr(Z = b \mid C)$, *conditional component means* $m_b^\pi(C) := \mathbb{E}[V_1 \mid Z = b, C]$, *and cross-mode diameter* $D_\pi(C) := \max_{b,b'} \|m_b^\pi(C) - m_{b'}^\pi(C)\|$. *Let* $c_\gamma(C) := \gamma_{b_-}^\pi(C)\gamma_{b_+}^\pi(C)$ *for a pair* (b_-, b_+) *attaining* $D_\pi(C)$. *Then*

$$\mathbb{E}\big[\|V_1 - V_0\|^2 \,\big|\, C\big]_{\mathrm{Gauss}} \geq c_\gamma(C) D_\pi^2(C) + d. \tag{10}$$

Theorem 3 gives an exact identity for the conditional training-target scale under the direct surrogate-source sampler, whose source marginal is exactly $\widetilde{\pi}(\cdot \mid C)$. The same conditioning argument applies to other designs of the label draw: whenever $V_0 \mid (B, C, V_1) \sim \mathcal{N}(m_B(C), \Lambda_B(t))$, one has $\mathbb{E}[\|V_1 - V_0\|^2 \mid C] = \mathbb{E}[\|V_1 - m_B(C)\|^2 + \operatorname{tr}\Lambda_B(t) \mid C]$, which reduces to (8) when B is drawn from $\gamma(C)$ independently of V_1 (Remark 20).

The total second-moment ratio is

$$\mathcal{R}_{\mathrm{tot}}(C) := \frac{\mathbb{E}[\|V_1 - V_0\|^2 \mid C]_{\mathrm{sur}}}{\mathbb{E}[\|V_1 - V_0\|^2 \mid C]_{\mathrm{Gauss}}} = \frac{S_\pi(C) + A_{\mathrm{sur}}(C) + B_{\mathrm{sur}}(C) + M_{\mathrm{sur}}(C)}{S_\pi(C) + \|\mu_\pi(C)\|^2 + d}.$$

Thus the direct surrogate source gives a smaller total training target than the independent Gaussian source whenever $A_{\mathrm{sur}}(C) + B_{\mathrm{sur}}(C) + M_{\mathrm{sur}}(C) < \|\mu_\pi(C)\|^2 + d$.

Equivalently, the surrogate source is beneficial when its within-component covariance, its weighted between-component spread, and its conditional mean mismatch are jointly smaller than the source penalty incurred by the isotropic Gaussian source.

It is useful to isolate the source-dependent part of the comparison. Since the intrinsic conditional spread $S_\pi(C)$ appears in both couplings, define the excess ratio

$$\mathcal{R}_{\mathrm{ex}}(C) := \frac{\mathbb{E}[\|V_1 - V_0\|^2 \mid C]_{\mathrm{sur}} - S_\pi(C)}{\mathbb{E}[\|V_1 - V_0\|^2 \mid C]_{\mathrm{Gauss}} - S_\pi(C)} = \frac{A_{\mathrm{sur}}(C) + B_{\mathrm{sur}}(C) + M_{\mathrm{sur}}(C)}{\|\mu_\pi(C)\|^2 + d}.$$

This excess ratio becomes small when the surrogate conditional source is concentrated near the true conditional velocity law:

$$A_{\mathrm{sur}}(C) \ll \|\mu_\pi(C)\|^2 + d, \qquad B_{\mathrm{sur}}(C) \ll \|\mu_\pi(C)\|^2 + d, \qquad M_{\mathrm{sur}}(C) \ll \|\mu_\pi(C)\|^2 + d.$$

In the extreme case $\widetilde{\rho}_1 = \rho_1$ at $t = 0$, the surrogate excess equals $\operatorname{tr}\mathrm{Cov}(X_1) \leq d$ under standardization or $[-1, 1]$ scaling, so it never exceeds the Gaussian excess (Remark 19). These statements concern the direct sampler; a Monte Carlo check on seven datasets matches both closed forms (Appendix §D).

5 Experiments

We evaluate CCVFM on four image datasets: three in raw pixel space, MNIST (28×28), CIFAR-10 ($32 \times 32 \times 3$), and ImageNet-32 [5] ($32 \times 32 \times 3$ with $1{,}000$ classes), all at matched U-Net architecture with HRF2, and one high-resolution latent-space benchmark, CelebA-HQ 256 ($256 \times 256 \times 3$, encoded by DC-AE f32c32 to a d=2048 latent), where the correction net is a DiT-L. All FID values use InceptionV3 pool features [12]; training details, hyperparameter sweeps, and toy 2D results are in the Appendix §C, D and E.

Table 1: MNIST pixel-space FID. CCVFM uses $K = 2000$, rank $r = 50$, U-Net base width 128, 200k training iterations, EMA 0.9999. FID_{50k} uses 50k training images as the reference (HRF2 Table 7 protocol); FID_{10k} uses the 10k held-out test set. RF and HRF2 baselines from Zhang et al. [31, Table 7]; DDPM and FM rows ($\star$) report 5k-reference numbers from their original papers for historical context.

Method	NFE	$\mathrm{FID}_{10k} \downarrow$	$\mathrm{FID}_{50k} \downarrow$
Rectified Flow [18]	10	—	7.974
Rectified Flow	100	—	5.563
HRF2 [31]	10	—	6.644
HRF2	100	—	2.588
HRF2	500	—	2.574
DDPM [13]*	1000	—	12.80
Flow Matching [17]*	100	—	10.20
CCVFM Stage III, L=5	6	13.80	12.45
CCVFM Stage III, L=10	11	4.05	2.88
CCVFM Stage III, L=20	21	**2.26**	**1.09**
CCVFM Stage III, L=50	51	**1.91**	**0.75**

FID protocol. We report two FID variants: FID_{50k} uses 50k generated samples against 50k *training* images (the protocol of 31, Table 7), and FID_{10k} uses 10k generated samples against the 10k held-out *test* set. Reported numbers are single-seed point estimates . On MNIST, five sampling seeds give $\mathrm{FID}_{50k} = 0.772 \pm 0.011$ at 51 NFE and a retrained model gives 0.747 (Appendix §D); the other benchmarks are single training runs.

5.1 MNIST: FID and goodness-of-fit diagnostics

Scaling law and HRF2 comparison. We trained three configurations with $K \in \{500, 1000, 2000\}$. At K=500 (40k iterations, U-Net base 64), CCVFM reaches $\mathrm{FID}_{50k} = 3.79$ at 11 NFE; at K=1000 (80k iterations, base 64), 1.93 at 11 NFE; at the headline K=2000 (200k iterations, base 128, rank 50), $\mathbf{1.09}$ at 21 NFE and $\mathbf{0.75}$ at 51 NFE (Table 1). The improvement in the best FID as K and the training budget grow is qualitatively consistent with the Stage-I compression assumption. Our L=20 result improves on HRF2's best (2.574 at 500 NFE) by 58% with $\approx 24\times$ fewer NFE; our L=50 result is $\approx 3\times$ lower at $\approx 10\times$ fewer NFE. Sample grids are in Figure 2 and Appendix §F.

Goodness of fit: the generator learns ρ_1, not the training set. A low FID alone cannot distinguish a generator that has learned the target law from one that has memorized its training set. To rule memorization out, we compare three pools of 10,000 samples each: a *Generated* pool (CCVFM with K=2000 and L=20, FID_{50k}=1.09), the 10,000-image held-out *Test* set, and a size-matched 10,000-image random subset of the *Training* set. For each ordered pair (A, B) we compute the nearest-neighbour distance $d_{A\to B}(a) = \min_{b\in B} \|\phi(a) - \phi(b)\|$ for each $a \in A$, where ϕ is the InceptionV3 pool embedding. Comparing the distributions of these distances across pairs is a direct, non-parametric goodness-of-fit probe; the absence of a memorization signal would manifest as symmetry (KS and Wasserstein-1 near zero) between the Generated-vs-Training and Generated-vs-Test comparisons.

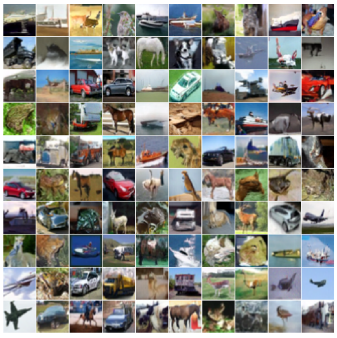
What the tables say. The memorization tests (Table 2, rows 3–4) give $\mathrm{KS} \le 0.024$, $W_1 \le 0.054$: Generated-to-Train and Generated-to-Test distance distributions are essentially the same. The real↔real baseline (row 5, $\mathrm{KS} = 0.008$, $W_1 = 0.012$) is the floor for two i.i.d. draws from ρ_1; the memorization test sits within a small factor of this floor, so the generated pool is distributed as a third i.i.d. draw, not preferentially closer to training. Improved precision/recall [16] computed in Inception feature space with k=5 are also indistinguishable across references ($\Delta P = 0.008$, $\Delta R = 0.004$; full table in Appendix §G). The three diagnostic panels (1-NN density, per-sample Generated-to-Train vs. Generated-to-Test scatter, improved precision/recall) are visualized in Appendix §G; pixel-space duplicates and CDF figures are also in Appendix §G.

Table 2: Goodness-of-fit summary on MNIST generations from the headline configuration (Inception feature space, size-matched pools of 10k samples, 1-nearest-neighbour). KS is the Kolmogorov–Smirnov statistic and W_1 the 1-Wasserstein distance between the two distance distributions named in "Pair A" and "Pair B". *Memorization tests* (rows 3, 4) compare generated-to-train against generated-to-test distances. The real↔real baseline (row 5) bounds how small these gaps can be even for two i.i.d. draws from ρ_1.

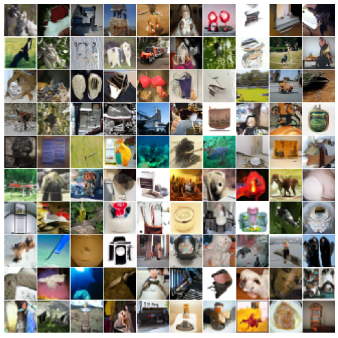
Pair A vs. Pair B	KS↓	$W_1 \downarrow$	Interpretation
Ge→Te vs. Te→Ge	0.0306	0.0728	fidelity symmetry (Ge vs. Test)
Ge→Tr vs. Tr→Ge	0.0248	0.0629	fidelity symmetry (Ge vs. Train)
Ge→Tr vs. Ge→Te	**0.0173**	**0.0420**	**no memorization signal**
Tr→Ge vs. Te→Ge	**0.0238**	**0.0535**	**no dual memorization**
Te→Tr vs. Tr→Te	0.0083	0.0116	real ↔ real baseline



(a) MNIST, 21 NFE, FID_{50k}=1.09



(b) CIFAR-10, 51 NFE, FID_{50k}=6.35



(c) ImageNet-32, 51 NFE, FID_{50k}=8.76



(d) CelebA-HQ 256, 51 NFE, FID_{28k}=4.17 (uncurated 2×5 panel)

Figure 2: Uncurated samples used for the reported FID pools: first 100 samples for (a)–(c), and a 2×5 CelebA-HQ panel from the seed-99 30k pool for (d).

5.2 High-dimensional images: CIFAR-10, ImageNet-32, and CelebA-HQ 256

We evaluate CCVFM on three pixel-/latent-space image benchmarks: CIFAR-10 (32×32×3, d=3072, n=50k), ImageNet-32 (32×32×3, d=3072, n≈1.28M), and CelebA-HQ 256 (256×256×3, encoded by DC-AE f32c32 to a d=2048 latent, n≈28k). Table 3 reports the three FID sweeps side by side.

On CIFAR-10 (Table 3a), CCVFM with K=10000 improves from $\text{FID}_{50k} = 9.66$ at 11 NFE to 6.35 at 51 NFE. ImageNet-32 (Table 3b) gives the cleanest HRF2 comparison (same U-Net, dataset, and protocol; only the source changes): CCVFM wins by 38% at 11 NFE and 24% at 21 NFE, and ties within 3% at 51 NFE. On CelebA-HQ 256 (Table 3c), the DC-AE f32c32 latent (d=2048) lets a K=10000 coreset capture face structure that DiT-B (131M params) under-fits; scaling to DiT-L (458M) reaches FID_{28k}=**4.17** at 51 NFE, competitive with ADM and LDM-4 at 4–20× fewer NFE without classifier-free guidance. Sample grids are in Figure 2 and Appendix §F. Ablations on L, K, the source/coupling choice, EMA, and the matched CIFAR-10 K=5000 → 10000 coreset-size sweep are in Appendix §D, together with Stage-I cost, rank, memorization, KDE-source, compression, and manifold studies, and the Stage-III training cost relative to HRF2.

Table 3: FID on 3 high-dimensional benchmarks. *(a)* CIFAR-10 FID_{50k} at $K{=}10000$, $r{=}80$; the matched $K{=}5000$ sweep is in Appendix §D; † denotes numbers not directly comparable (different protocol). *(b)* ImageNet-32 FID_{50k} (matched U-Net, same protocol; HRF2 from Zhang et al. [31]). *(c)* CelebA-HQ 256 FID_{28k} (28k generated vs. all $\sim$28k real CelebA-HQ images) at $K{=}10000$, $r{=}128$, DiT-L correction ($\sim$458M params), 400k training steps, no classifier-free guidance.

Method	NFE	FID_{50k}
RF [18]†	1	378.0
2-RF [18]†	1	12.21
DDPM [13]	1000	3.17
FM [17]	adp.	6.35
EDM [14]	35	1.97
CM [26]	1	8.70
CCVFM, $L{=}10$	11	9.66
CCVFM, $L{=}20$	21	7.24
CCVFM, $L{=}50$	51	**6.35**

(a) CIFAR-10.

Method	NFE	FID_{50k}	Δ
HRF2	11	20.29	—
CCVFM, $L{=}10$	11	**12.55**	-38%
HRF2	21	12.49	—
CCVFM, $L{=}20$	21	**9.51**	-24%
HRF2	51	9.02	—
CCVFM, $L{=}50$	51	**8.76**	-3%

(b) ImageNet-32. Δ vs. HRF2.

Method	NFE	FID_{28k}
DDPM [13]	1000	7.03
LSGM	138	7.22
ADM	1000	5.11
LDM-4	200	5.11
EDM [14]	79	3.50
CCVFM-B, $L{=}50$	51	5.59
CCVFM-L, $L{=}50$	51	**4.17**

(c) CelebA-HQ 256.

6 Related Work

Generative modeling and few-step sampling. GANs [9], VAEs [15], and normalizing flows [22, 8] established the modern toolkit, with tradeoffs in stability, likelihood bounds, or architecture. Score-based diffusion models [13, 25, 14] drive recent image quality but often need long chains; EDM [14] is the strongest CIFAR-10 efficiency benchmark. Consistency models [26] and progressive distillation [24] compress trained teachers, while ReFlow [18] and mini-batch OT-CFM [27, 21] straighten paths. CCVFM is orthogonal: it reduces the inner correction transport before any distillation, straightening, OT coupling, or solver choice.

Flow matching and velocity distributions. Conditional flow matching [17], rectified flow [18], and stochastic interpolants [1] regress the conditional mean velocity, which is the squared-loss optimum but collapses multimodal local velocity laws. The closest prior work is HRF2 [31], which models $\pi(v \mid x_t, t)$ via a second-level velocity-space flow. We keep HRF2's identity and closed-form GMM corollary, but replace its $\mathcal{N}(0, I)$ source with a coreset-GMM surrogate whose conditional law is sampled in closed form through its mixture representation. Concurrently, Guo and Schwing [11] proposed a variational low-dimensional latent route to velocity multimodality.

Coresets, smoothing, and diagnostics. Stage I builds on entropic Optimal Transport (OT) [7] and weighted coreset / quantization theory [6, 10, 30]. Classical quantization provides the optimal K-atomic benchmark, while the deployed Sinkhorn / GMM rate is used through the explicit Stage-I compression assumption. In our diagnostic, we use improved precision / recall [23, 16, 20], and a 1-NN KS/W_1 memorization test built on prior nearest-neighbour-based generative-model diagnostics [2, 19, 3, 4] .

7 Conclusion and Limitations

CCVFM replaces the velocity-space flow's isotropic noise source with a closed-form coreset-GMM surrogate and trains the correction network from that surrogate source. The surrogate gap equals $W_2(\rho_1, \widetilde{\rho}_1)$ and vanishes with K under the Stage-I compression assumption, while the noise-source analogue has a dimension-scale lower bound that is constant in K and n (Theorem 1). Under the direct surrogate-source sampler, the source-dependent training excess is small whenever the surrogate is close to π in mean and covariance (Theorem 3).

Limitations include pessimistic high-d constants, the assumed compression rate for the deployed coreset/GMM, residual-field regularity in the NFE scaling, EDM's remaining CIFAR-10 lead at matched NFE, and single-run results on CIFAR-10, ImageNet-32, and CelebA-HQ. Future work should prove algorithmic rates for the entropic-Sinkhorn/GMM stage, extend the analysis to other target-density classes and to $J > 1$, extend CCVFM to text-to-image generation via per-condition coresets, and combine it with distillation/straightening.

Acknowledgements and Disclosures

Xiao Wang acknowledges support from the National Science Foundation under Grant No. SES-2316428 and from Anthropic through its AI for Science program.

References

[1] Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

[2] Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *IEEE Symposium on Security and Privacy (S&P)*, 2022.

[3] Ali Borji. Pros and cons of GAN evaluation measures. In *Computer Vision and Image Understanding*, 2019.

[4] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, 2023.

[5] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of ImageNet as an alternative to the CIFAR datasets. In *arXiv preprint arXiv:1707.08819*, 2017.

[6] Sebastian Claici, Aude Genevay, and Justin Solomon. Wasserstein measure coresets. *arXiv preprint arXiv:1805.07412*, 2018.

[7] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 2013.

[8] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[9] Ian J. 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.

[10] Siegfried Graf and Harald Luschgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer, 2000.

[11] Pengsheng Guo and Alexander G. Schwing. Variational rectified flow matching. In *International Conference on Machine Learning*, 2025.

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[14] 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.

[15] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[16] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, 2019.

[17] 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.

[18] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

[19] Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta. A non-parametric test to detect data-copying in generative models. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[20] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, 2020.

[21] Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Ricky T. Q. Chen, and Yaron Lipman. Multisample flow matching: Straightening flows with minibatch couplings. *arXiv preprint arXiv:2304.14772*, 2023.

[22] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[23] Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, 2018.

[24] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

[25] 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.

[26] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.
[27] Alexander Tong, Nikolay Malkin, Kilian Fatras, Lazar Atanackovic, Yan Zhang, Guillaume Huguet, Guy Wolf, Yoshua Bengio, and Aaron Courville. Improving and generalizing flow-based generative models with minibatch optimal transport. In *Transactions on Machine Learning Research*, 2024.
[28] Cédric Villani. Optimal transport: Old and new. *Grundlehren der mathematischen Wissenschaften*, 338, 2009.
[29] Xiao Wang, Zihua She, and Jianxi Su. Optimal estimation for flow matching via coreset induced coupling. Submitted to *The Annals of Statistics*, 2026.
[30] Haoyun Yin, Yixuan Qiu, and Xiao Wang. Wasserstein coreset via sinkhorn loss. *Transactions on Machine Learning Research*, 2025. URL `https://openreview.net/forum?id=DrMCDS88IL`.
[31] Yichi Zhang, Yici Yan, Alex Schwing, and Zhizhen Zhao. Towards hierarchical rectified flow. In *International Conference on Learning Representations*, 2025.

A Proofs

A.1 Notation

Throughout, $X_0 \sim \rho_0 = \mathcal{N}(0, I_d)$ and $X_1 \sim \rho_1$ are independent, $X_t = (1-t)X_0 + tX_1$, $V = X_1 - X_0$. The true conditional velocity law is $\pi(v \mid x, t)$; the surrogate is $\widetilde{\pi}(v \mid x, t)$ induced by replacing ρ_1 with the coreset GMM $\widetilde{\rho}_1$. We identify a density with its induced probability measure whenever no ambiguity is possible; if the Stage-I surrogate is a finite coreset, the same notation $\widetilde{\rho}_1$ denotes the corresponding probability measure. For a surrogate conditional source $Q(\cdot \mid x, t)$, the *conditional transport cost* (Eq. (6) of the main paper) is

$$D(Q) \;=\; \Big(\mathbb{E}_{X_0\sim\rho_0}\big[W_2^2\big(Q(\cdot \mid X_0, 0),\, \pi(\cdot \mid X_0, 0)\big)\big]\Big)^{1/2}, \tag{11}$$

defined in the second-moment form, consistent with the $D(Q)$ used in Theorem 6.

Table 4: Main notation.

Symbol	Meaning
$\rho_0 = \mathcal{N}(0, I_d)$, ρ_1	outer source and data law
$X_t = (1-t)X_0 + tX_1$, $V = X_1 - X_0$	outer interpolation and velocity
$C = (X_t, t)$	conditioning state of the inner flow
$\pi(\cdot \mid C)$, $\widetilde{\pi}(\cdot \mid C)$	true and surrogate conditional velocity laws
μ_π, Σ_π; $\mu_{\mathrm{sur}}, \Sigma_{\mathrm{sur}}$	conditional means / covariances of π, $\widetilde{\pi}$
$\widetilde{\rho}_1 = \sum_k w_k \mathcal{N}(\mu_k, \Sigma_k)$	Stage-I coreset-GMM surrogate, $\Sigma_k = L_k L_k^\top + \sigma_k^2 I_d$
K, r	number of components, covariance rank
$\gamma_b(C)$, $m_b(C)$, $\Lambda_b(t)$	weights, means, covariances of $\widetilde{\pi}(\cdot \mid C)$
V_0, V_1, τ, V_τ	inner source, inner target, inner time, inner interpolant
$f_\theta(v, \tau, x_t, t)$	Stage-III correction network
$D(Q)$	conditional transport cost (11)
J, L	outer steps, inner Euler steps (NFE $= J(L+1)$)

A.2 Conditional velocity identity and GMM specialization

We restate these for completeness; the proofs are standard (see [31]).

Proposition 4 (Conditional velocity identity, § 2 main paper, Eq. (1))**.** *For any $t \in (0,1)$ and x with $\rho_t(x) > 0$,*

$$\pi(v \mid x, t) = \frac{\rho_0(x - tv)\, \rho_1(x + (1-t)v)}{\rho_t(x)}.$$

At $t = 0$, by independence of X_0, X_1, $\pi(v \mid x_0, 0) = \rho_1(x_0 + v)$.

Proof. The joint density of (X_t, V) at (x, v) is $\rho_0(x - tv)\rho_1(x + (1-t)v)$ by independence of the endpoints combined with the endpoint identities $X_0 = X_t - tV$, $X_1 = X_t + (1-t)V$. Normalize by $\rho_t(x)$. At $t = 0$, $X_t = X_0$ so $V = X_1 - x_0$ with density $\rho_1(v + x_0)$. □

Proposition 5 (GMM specialization, Prop. 4 with $\rho_1 = \widetilde{\rho}_1$)**.** *If $\widetilde{\rho}_1 = \sum_b w_b \mathcal{N}(\mu_b, \Sigma_b)$, then the induced $\widetilde{\pi}(v \mid x, t) = \sum_b \gamma_b(x,t)\,\mathcal{N}(v; m_b(x,t), \Lambda_b(t))$ is a K-component Gaussian mixture with closed-form parameters:*

$$\begin{aligned} A_b(t) &= t^2 I + (1-t)^2 \Sigma_b^{-1}, \quad \Lambda_b(t) = A_b(t)^{-1}, \\ b_b(x,t) &= tx - (1-t)\Sigma_b^{-1}(x - \mu_b), \quad m_b(x,t) = \Lambda_b(t) b_b(x,t), \\ \gamma_b(x,t) &\propto w_b\, \eta_b(x,t), \quad \eta_b = |\Lambda_b|^{1/2} |\Sigma_b|^{-1/2} e^{\frac{1}{2} b_b^\top \Lambda_b b_b - \frac{1}{2} c_b(x)}, \\ c_b(x) &= x^\top x + (x-\mu_b)^\top \Sigma_b^{-1}(x - \mu_b). \end{aligned}$$

At $t = 0$: $\widetilde{\pi}(v \mid x_0, 0) = \widetilde{\rho}_1(x_0 + v) = \sum_b w_b \mathcal{N}(v; \mu_b - x_0, \Sigma_b)$.

Proof. Completing the square in v in $\mathcal{N}(x - tv; 0, I)\mathcal{N}(x + (1-t)v; \mu_b, \Sigma_b)$ yields a Gaussian in v with natural parameters A_b and b_b, hence mean $m_b = \Lambda_b b_b$ and covariance Λ_b. The multiplicative constant for each component is collected into η_b, and normalization produces γ_b. The $t = 0$ case is immediate from $\pi(v \mid x_0, 0) = \rho_1(x_0 + v)$ in Prop. 4 with $\rho_1 \mapsto \widetilde{\rho}_1$. □

A.3 Theorem 1 of main paper (Transport-task reduction)

Theorem 6. *With $D(Q)$ as in* (11)*, define $Q_{\mathrm{CCVFM}}(\cdot \mid x_0, 0) = (S_{x_0})_{\#}\widetilde{\rho}_1$ with $S_{x_0}(y) = y - x_0$ (equivalently, when $\widetilde{\rho}_1$ has a density, $Q_{\mathrm{CCVFM}}(v \mid x_0, 0) = \widetilde{\rho}_1(x_0 + v)$), and $Q_{\mathrm{HRF2}}(v \mid x_0, 0) = \mathcal{N}(v; 0, I_d)$. Assume $\rho_0 = \mathcal{N}(0, I_d)$ and $X_0 \perp X_1$ with $X_1 \sim \rho_1 \in \mathcal{P}_2(\mathbb{R}^d)$. Then*

(i) $D(Q_{\mathrm{CCVFM}}) = W_2(\rho_1, \widetilde{\rho}_1)$*;*

(ii) $D(Q_{\mathrm{HRF2}}) \geq \sqrt{d}\,\big|\sqrt{\sigma_1^2 + 1} - 1\big|$*, where* $\sigma_1^2 := \mathbb{E}\|X_1\|^2/d$.

Proof. **(i)** By Prop. 4, at $t = 0$ the true conditional velocity law is $\pi(\cdot \mid x_0, 0) = \mathcal{L}(X_1 - x_0) = (S_{x_0})_{\#}\rho_1$. The CCVFM source is $(S_{x_0})_{\#}\widetilde{\rho}_1$. For any coupling (Y_1, Y_2) of ρ_1 and $\widetilde{\rho}_1$, the shifted pair $(Y_1 - x_0, Y_2 - x_0)$ couples the two velocity laws, and conversely any velocity-space coupling can be shifted by x_0 to couple ρ_1 and $\widetilde{\rho}_1$. The quadratic cost is unchanged under this common shift, so for every fixed x_0,

$$W_2\big(Q_{\mathrm{CCVFM}}(\cdot \mid x_0, 0),\, \pi(\cdot \mid x_0, 0)\big) \;=\; W_2(\widetilde{\rho}_1, \rho_1), \tag{12}$$

which is constant in x_0. Hence $\mathbb{E}_{X_0} W_2^2 = W_2(\widetilde{\rho}_1, \rho_1)^2$, and $D(Q_{\mathrm{CCVFM}}) = W_2(\widetilde{\rho}_1, \rho_1)$.

(ii) We proceed in four labeled steps.

Step 1 (W_2^2 moment lower bound). For any probability measures μ, ν on $\mathbb{R}^d$ with finite second moments,

$$W_2^2(\mu, \nu) \;\geq\; \big(\sqrt{m_2(\mu)} - \sqrt{m_2(\nu)}\big)^2, \tag{13}$$

where $m_2(\mu) := \int \|x\|^2 \,\mathrm{d}\mu(x)$. This is immediate from any coupling (U, V): by the Minkowski inequality, $(\mathbb{E}\|U - V\|^2)^{1/2} \geq |(\mathbb{E}\|U\|^2)^{1/2} - (\mathbb{E}\|V\|^2)^{1/2}|$. Taking the infimum over couplings yields (13) (see, e.g., [28]).

Step 2 (Apply (13) *at fixed x_0).* Fix $x_0 \in \mathbb{R}^d$. We have $m_2(\mathcal{N}(0, I_d)) = d$. For $\mu = \mathcal{L}(X_1 - x_0)$ with $X_1 \sim \rho_1$,

$$m_2(\mu) \;=\; \mathbb{E}_{X_1}\|X_1 - x_0\|^2 \;=:\; S(x_0)^2,$$

a deterministic quantity depending only on x_0. By (13),

$$W_2^2\big(\mathcal{N}(0, I_d),\, \mathcal{L}(X_1 - x_0)\big) \;\geq\; \big(\sqrt{d} - S(x_0)\big)^2. \tag{14}$$

Step 3 (Take expectation over $X_0 \sim \rho_0$). Integrating (14) over X_0 and expanding the square,

$$\mathbb{E}_{X_0}\big[(\sqrt{d} - S(X_0))^2\big] \;=\; d \;-\; 2\sqrt{d}\,\mathbb{E}_{X_0}[S(X_0)] \;+\; \mathbb{E}_{X_0}[S(X_0)^2]. \tag{15}$$

We compute each moment of S:

$$\begin{aligned} \mathbb{E}_{X_0}[S(X_0)^2] &= \mathbb{E}_{X_0}\mathbb{E}_{X_1}\|X_1 - X_0\|^2 = \mathbb{E}\|X_1\|^2 \;-\; 2\,\mathbb{E}\langle X_1, X_0\rangle \;+\; \mathbb{E}\|X_0\|^2 \\ &= d\sigma_1^2 \;-\; 2\,\langle \mu_{\rho_1}, \mathbb{E}X_0\rangle \;+\; d \;=\; d\sigma_1^2 + d \;=\; d(\sigma_1^2 + 1), \end{aligned} \tag{16}$$

using $X_0 \perp X_1$, $\mu_{\rho_0} = \mathbb{E}X_0 = 0$, and $\mathbb{E}\|X_0\|^2 = d$ (standard Gaussian). By Jensen's inequality with concave $u \mapsto \sqrt{u}$,

$$\mathbb{E}_{X_0}[S(X_0)] \;=\; \mathbb{E}_{X_0}\sqrt{S(X_0)^2} \;\le\; \sqrt{\mathbb{E}_{X_0}[S(X_0)^2]} \;=\; \sqrt{d(\sigma_1^2+1)}. \tag{17}$$

The coefficient $-2\sqrt{d}$ in front of $\mathbb{E}S$ in (15) is *negative*, so substituting the *upper bound* (17) produces a valid *lower bound* on the right-hand side:

$$\begin{aligned}\mathbb{E}_{X_0}[(\sqrt{d} - S(X_0))^2] &\ge\; d \;-\; 2\sqrt{d}\cdot\sqrt{d(\sigma_1^2+1)} \;+\; d(\sigma_1^2+1)\\ &=\; d\left[1 \;-\; 2\sqrt{\sigma_1^2+1} \;+\; (\sigma_1^2+1)\right]\\ &=\; d\left(\sqrt{\sigma_1^2+1}-1\right)^2,\end{aligned} \tag{18}$$

where the last line uses the algebraic identity $u^2 - 2u + 1 = (u-1)^2$ with $u = \sqrt{\sigma_1^2+1}$.

Step 4 (Collect). Combining (14) and (18) (and noting that both inequalities preserve their direction after integration),

$$D(Q_{\mathrm{HRF2}})^2 \;=\; \mathbb{E}_{X_0}\big[W_2^2\big(\mathcal{N}(0, I_d),\, \mathcal{L}(X_1 - X_0)\big)\big] \;\ge\; d\left(\sqrt{\sigma_1^2+1}-1\right)^2. \tag{19}$$

Taking the non-negative square root and using $|a| = \sqrt{a^2}$ gives

$$D(Q_{\mathrm{HRF2}}) \;\ge\; \sqrt{d}\,\big|\sqrt{\sigma_1^2+1}-1\big|, \tag{20}$$

which is the statement of (ii). □

Remark 7 (Why the bound is nonzero and K-independent). Since $\sigma_1^2 \ge 0$, $\sqrt{\sigma_1^2+1} \ge 1$, with equality iff $\sigma_1 = 0$ (i.e. $\rho_1 = \delta_0$). For every nondegenerate target the bound is strictly positive and, crucially, depends only on the second moment of ρ_1, not on the coreset size K or sample size n. Typical image targets have σ_1 bounded and d large, so the bound scales as $\Omega(\sqrt{d}\,\sigma_1^2)$ for small σ_1 and as $\Omega(\sqrt{d}\,\sigma_1)$ for large σ_1.

Lemma 8 (Existence of a K-atom quantization benchmark)**.** *Assume ρ_1 is supported on $\Omega = [0,1]^d$. Let*

$$\mathcal{G}_K(\Omega) := \left\{ \sum_{j=1}^K w_j \delta_{y_j} :\; w_j \ge 0,\; \sum_{j=1}^K w_j = 1,\; y_j \in \Omega \right\}.$$

Then there exists $C_d < \infty$, depending only on d, such that

$$\inf_{\nu \in \mathcal{G}_K(\Omega)} W_2(\rho_1, \nu) \le C_d K^{-1/d}.$$

Consequently, if the Stage-I coreset $\widetilde{\rho}_{1,K}$ is a global minimizer of the K-atom Wasserstein quantization problem over $\mathcal{G}_K(\Omega)$, then

$$W_2(\rho_1, \widetilde{\rho}_{1,K}) = O(K^{-1/d}).$$

Proof. Let $m = \lfloor K^{1/d} \rfloor$. For K large enough, $m \ge c_d K^{1/d}$ for a constant $c_d > 0$ and $m^d \le K$. Partition $[0,1]^d$ into m^d axis-aligned cubes $\{A_\ell\}$ of side length $1/m$. For each cell with $\rho_1(A_\ell) > 0$, choose $a_\ell \in A_\ell$ and set

$$\nu_m := \sum_\ell \rho_1(A_\ell)\delta_{a_\ell}.$$

This measure has at most $m^d \le K$ atoms. Couple $X \sim \rho_1$ to a_ℓ whenever $X \in A_\ell$. Since every cube has diameter at most $\sqrt{d}/m$,

$$W_2^2(\rho_1, \nu_m) \le \sum_\ell \int_{A_\ell} \|x - a_\ell\|^2 \,\mathrm{d}\rho_1(x) \le \frac{d}{m^2}.$$

Thus $W_2(\rho_1, \nu_m) \le \sqrt{d}/m \le C_d K^{-1/d}$. Enlarging C_d covers the finitely many small values of K. Taking the infimum over $\mathcal{G}_K(\Omega)$ proves the lemma. □

Consequence for CCVFM. If ρ_1 is supported on $[0,1]^d$ and the Stage-I coreset $\widetilde{\rho}_{1,K}$ is a global K-atom Wasserstein quantizer, then

$$D(Q_{\mathrm{CCVFM}}) = W_2(\rho_1, \widetilde{\rho}_{1,K}) = O(K^{-1/d})$$

by Theorem 6(i) and Lemma 8. This consequence applies to exact global Wasserstein quantization. It does not assert that the finite-iteration entropic-Sinkhorn/GMM routine used in Stage I necessarily returns such a global quantizer. For the deployed routine, the corresponding approximation property is imposed as Assumption 1 of the main paper.

Remark 9 (How to read the coreset assumption). The preceding consequence uses only the W_2 approximation property $W_2(\rho_1, \widetilde{\rho}_{1,K}) = O(K^{-1/d})$. If Stage I is not solved as an exact global quantization problem, it is enough to assume or prove that the returned coreset satisfies the same rate, possibly with additional optimization, entropic-regularization, and GMM-lift slack. This paper treats that deployed-rate statement as the Stage-I compression assumption rather than as a consequence of the algorithm.

A.4 Corollary 2 of main paper (NFE sufficient scaling)

Lemma 10 (Euler error under space-time regularity)**.** *Consider the non-autonomous ODE*

$$\dot{V}_\tau = f_\theta(V_\tau, \tau), \qquad \tau \in [0,1].$$

Assume f_θ is L_v-Lipschitz in v and L_τ-Lipschitz in τ, and assume along the exact trajectory

$$\sup_{\tau\in[0,1]} \|f_\theta(V_\tau, \tau)\|_{L^2} \le M.$$

Then the explicit Euler scheme with n uniform steps satisfies

$$\|\mathrm{Euler}(n) - \mathrm{Exact}\|_{L^2} \le \frac{e^{L_v} - 1}{2L_v} \frac{L_v M + L_\tau}{n}, \tag{21}$$

where the prefactor is interpreted as $1/2$ when $L_v = 0$. In particular, if L_v is bounded, then

$$\|\mathrm{Euler}(n) - \mathrm{Exact}\|_{L^2} = O\left(\frac{L_v M + L_\tau}{n}\right).$$

Proof. Let $h = 1/n$ and $\tau_k = kh$. Denote by V_{τ_k} the exact solution and by $\widetilde{V}_k$ the Euler iterate. The one-step local truncation error is

$$\delta_k := V_{\tau_{k+1}} - V_{\tau_k} - h f_\theta(V_{\tau_k}, \tau_k) = \int_{\tau_k}^{\tau_{k+1}} \big(f_\theta(V_t, t) - f_\theta(V_{\tau_k}, \tau_k)\big)\, \mathrm{d}t.$$

Using the Lipschitz assumptions and the bound by M,

$$\|\delta_k\|_{L^2} \le \int_{\tau_k}^{\tau_{k+1}} (L_v M + L_\tau)(t - \tau_k)\, \mathrm{d}t = \frac{1}{2}(L_v M + L_\tau) h^2.$$

For the global error $e_k := \|\widetilde{V}_k - V_{\tau_k}\|_{L^2}$,

$$e_{k+1} \le (1 + hL_v) e_k + \frac{1}{2}(L_v M + L_\tau) h^2.$$

Iterating this recursion with $e_0 = 0$ gives

$$e_n \le \frac{(1+hL_v)^n - 1}{hL_v} \cdot \frac{1}{2}(L_v M + L_\tau) h^2 \le \frac{e^{L_v} - 1}{2L_v} \frac{L_v M + L_\tau}{n}.$$

The case $L_v = 0$ follows by taking the limit. □

Corollary 11 (NFE sufficient scaling)**.** *Let $D := D(Q)$ be the conditional transport cost in* (11)*. Assume the correction ODE satisfies the regularity conditions of Lemma 10. Assume further the residual straight-line scale and flattening conditions*

$$M = O(D), \qquad L_v = O(D), \qquad L_\tau = O(D^2), \tag{22}$$

with constants uniform over the family of transport tasks under consideration. Then, for any $\varepsilon_{\mathrm{disc}} > 0$, a sufficient number of Euler function evaluations to make the L^2 discretization error at most $\varepsilon_{\mathrm{disc}}$ is

$$\mathrm{NFE}(\varepsilon_{\mathrm{disc}}) = O\left(1 + \frac{D^2}{\varepsilon_{\mathrm{disc}}}\right). \tag{23}$$

Without the flattening conditions (22)*, the generic sufficient bound is*

$$\mathrm{NFE}(\varepsilon_{\mathrm{disc}}) = O\left(1 + \frac{L_v D + L_\tau}{\varepsilon_{\mathrm{disc}}}\right), \tag{24}$$

assuming only $M = O(D)$ and bounded L_v.

For CCVFM, let $\Delta_K := D(Q_{\mathrm{CCVFM}}) = W_2(\rho_1, \widetilde{\rho}_1)$. Then

$$\mathrm{NFE}_{\mathrm{CCVFM}}(\varepsilon_{\mathrm{disc}}) = O\left(1 + \frac{\Delta_K^2}{\varepsilon_{\mathrm{disc}}}\right)$$

under (22)*. If the Stage-I compression assumption gives $\Delta_K = O(K^{-1/d})$, then*

$$\mathrm{NFE}_{\mathrm{CCVFM}}(\varepsilon_{\mathrm{disc}}) = O\left(1 + \frac{K^{-2/d}}{\varepsilon_{\mathrm{disc}}}\right).$$

For HRF2, if d and ρ_1 are fixed and the corresponding regularity constants are $O(1)$, then a sufficient scaling is

$$\mathrm{NFE}_{\mathrm{HRF2}}(\varepsilon_{\mathrm{disc}}) = O\left(1 + \frac{1}{\varepsilon_{\mathrm{disc}}}\right).$$

Proof. Lemma 10 gives Euler error $O((L_v M + L_\tau)/n)$ when L_v is bounded. Under (22), $L_v M + L_\tau = O(D^2)$, so $n \geq C D^2/\varepsilon_{\mathrm{disc}}$ is sufficient; at least one function evaluation gives (23). If only $M = O(D)$ is used, the same Euler bound gives the generic scaling (24).

For CCVFM, $D(Q_{\mathrm{CCVFM}}) = \Delta_K$ by Theorem 6(i). This gives the displayed $\Delta_K^2/\varepsilon_{\mathrm{disc}}$ scaling. If the Stage-I compression assumption gives $\Delta_K = O(K^{-1/d})$, then the $K^{-2/d}/\varepsilon_{\mathrm{disc}}$ scaling follows. For HRF2, the law Q_{HRF2} is independent of K; for fixed d and a fixed nondegenerate target, Theorem 6(ii) gives a positive K-independent lower bound. A matching K-independent upper bound follows by the independent coupling: for $Z \sim \mathcal{N}(0, I_d)$ independent of (X_0, X_1),

$$D(Q_{\mathrm{HRF2}})^2 \leq \mathbb{E}\|Z - X_1 + X_0\|^2 = d(\sigma_1^2 + 2).$$

Thus $D(Q_{\mathrm{HRF2}}) = \Theta_K(1)$, and the sufficient NFE upper bound is $O(1 + 1/\varepsilon_{\mathrm{disc}})$ under $O(1)$ regularity constants. □

Remark 12 (Interpretation of the flattening assumption)*.* A small conditional transport cost $D(Q)$ controls the typical displacement magnitude, but it does not by itself determine the v-Lipschitz constant of the learned velocity field. The scaling $L_v = O(D)$ should therefore be read as an additional regularity or flattening assumption on the residual correction field. A sufficient regime is that, on a fixed-scale velocity domain,

$$f_{\theta,Q}(v, \tau) = D(Q)\, \widetilde{f}_{\theta,Q}(v, \tau), \qquad \mathrm{Lip}_v(\widetilde{f}_{\theta,Q}) = O(1).$$

For the non-autonomous ODE one also needs the time variation to be controlled, as in $L_\tau = O(D^2)$, or an equivalent local truncation-error assumption. Under these additional conditions the Euler upper bound scales as $O(D(Q)^2/n)$. Without them, the proof only yields the generic sufficient bound in (24), with the regularity constants treated as problem-dependent quantities.

The three conditions in (22) are assumptions on the learned residual field, not consequences of a small D alone. A useful reference is the boundary case $D = 0$: the source and target conditional laws then coincide, the identity correction $V_\tau \equiv V_0$ is valid, and for this reference solution $f_\theta \equiv 0$, so $M = L_v = L_\tau = 0$. The conditions specify how the learned field is assumed to approach the identity correction as D decreases. (i) $M = O(D)$ is the amplitude condition: since $V_1 - V_0 = \int_0^1 f_\theta(V_\tau, \tau)\, \mathrm{d}\tau$, M bounds the L^2 amplitude of the correction speed, uniformly in τ. (ii) $L_v = O(D)$ is the spatial flattening condition: it excludes a correction field that is small but varies sharply with v, and gives a spatial contribution $L_v M = O(D^2)$. (iii) $L_\tau = O(D^2)$ is the temporal regularity condition, placing the explicit time variation at the same order. Then $L_v M + L_\tau = O(D^2)$, and the Euler bound converts D into the D^2-NFE scaling. Corollary 11 is therefore a conditional sufficient result; without these conditions the general bound (21) still holds, and retaining only $M = O(D)$ and bounded L_v gives the weaker scaling (24).

A.5 Theorem 3 of main paper (Direct surrogate-source second moment vs. independent Gaussian source)

This subsection formalizes the training-signal scale for the direct surrogate-source sampler defined in Stage III. The relevant facts are: (i) the sampler has the exact source marginal $\widetilde{\pi}(\cdot \mid C)$, and (ii) the conditional second moment of the regression target admits a closed-form identity in terms of the conditional means and covariances of $\pi(\cdot \mid C)$ and $\widetilde{\pi}(\cdot \mid C)$.

Remark 13 (Second moment versus centered variance). The statements below control $\mathbb{E}[\|V_1 - V_0\|^2 \mid C]$, not the centered variance

$$\mathbb{E}\big[\|V_1 - V_0 - \mathbb{E}[V_1 - V_0 \mid C]\|^2 \,\big|\, C\big].$$

This is intentional: the squared norm of the training target $V_1 - V_0$ is the scale appearing directly in the flow-matching regression loss. Since centered variance is always bounded above by the second moment, the non-centered second moment is the cleaner quantity for training-signal size.

Proposition 14 (Direct surrogate sampling: correct marginal and conditional second moment)**.** *Fix $C = (X_t, t)$. Let $V_1 \sim \pi(\cdot \mid C)$ have finite second moment. Let the surrogate mixture be*

$$\widetilde{\pi}(v \mid C) = \sum_{b=1}^{K} \gamma_b(C)\, \mathcal{N}\big(v; m_b(C), \Lambda_b(t)\big).$$

Draw

$$B \mid C \sim \mathrm{Cat}\big(\gamma_1(C), \dots, \gamma_K(C)\big), \qquad V_0 \mid (B = b, C) \sim \mathcal{N}\big(m_b(C), \Lambda_b(t)\big),$$

and assume that this draw is conditionally independent of V_1 given C. Define

$$\mu_\pi(C) := \mathbb{E}[V_1 \mid C], \qquad \Sigma_\pi(C) := \mathrm{Cov}(V_1 \mid C),$$

and

$$\mu_{\mathrm{sur}}(C) := \sum_{b=1}^{K} \gamma_b(C) m_b(C),$$

$$\Sigma_{\mathrm{sur}}(C) := \sum_{b=1}^{K} \gamma_b(C) \Big[\Lambda_b(t) + \big(m_b(C) - \mu_{\mathrm{sur}}(C)\big)\big(m_b(C) - \mu_{\mathrm{sur}}(C)\big)^\top\Big].$$

Then

$$V_0 \mid C \sim \widetilde{\pi}(\cdot \mid C),$$

and

$$\mathbb{E}\big[\|V_1 - V_0\|^2 \,\big|\, C\big] = \mathrm{tr}\, \Sigma_\pi(C) + \mathrm{tr}\, \Sigma_{\mathrm{sur}}(C) + \big\|\mu_\pi(C) - \mu_{\mathrm{sur}}(C)\big\|^2. \tag{25}$$

Moreover, if

$$T_{\max}(C) := \max_{1 \le b \le K} \mathrm{tr}\, \Lambda_b(t), \qquad D_\mu(C) := \max_{1 \le b, b' \le K} \|m_b(C) - m_{b'}(C)\|,$$

then

$$\mathrm{tr}\, \Sigma_{\mathrm{sur}}(C) \le T_{\max}(C) + \frac{1}{2} D_\mu^2(C), \tag{26}$$

and therefore

$$\mathbb{E}\big[\|V_1 - V_0\|^2 \,\big|\, C\big] \le \mathrm{tr}\, \Sigma_\pi(C) + \big\|\mu_\pi(C) - \mu_{\mathrm{sur}}(C)\big\|^2 + T_{\max}(C) + \frac{1}{2} D_\mu^2(C). \tag{27}$$

Proof. Throughout the proof, all expectations are conditional on the fixed value of C. Write

$$\gamma_b = \gamma_b(C), \qquad m_b = m_b(C), \qquad \Lambda_b = \Lambda_b(t).$$

First, the marginal law of V_0 is obtained by marginalizing over the component label:

$$p(V_0 = v \mid C) = \sum_{b=1}^{K} \Pr(B = b \mid C)\, p(V_0 = v \mid B = b, C).$$

Substituting the sampling rule gives

$$p(V_0 = v \mid C) = \sum_{b=1}^{K} \gamma_b \mathcal{N}(v; m_b, \Lambda_b) = \widetilde{\pi}(v \mid C).$$

Hence $V_0 \mid C \sim \widetilde{\pi}(\cdot \mid C)$.

Next, the conditional mean of V_0 is

$$\mathbb{E}[V_0 \mid C] = \sum_{b=1}^{K} \gamma_b m_b = \mu_{\mathrm{sur}}(C).$$

By the law of total covariance,

$$\mathrm{Cov}(V_0 \mid C) = \mathbb{E}\big[\mathrm{Cov}(V_0 \mid B, C) \mid C\big] + \mathrm{Cov}\big(\mathbb{E}[V_0 \mid B, C] \mid C\big).$$

Since

$$\mathrm{Cov}(V_0 \mid B = b, C) = \Lambda_b, \qquad \mathbb{E}[V_0 \mid B = b, C] = m_b,$$

we obtain

$$\mathrm{Cov}(V_0 \mid C) = \sum_{b=1}^{K} \gamma_b \Lambda_b + \sum_{b=1}^{K} \gamma_b (m_b - \mu_{\mathrm{sur}}(C))(m_b - \mu_{\mathrm{sur}}(C))^\top = \Sigma_{\mathrm{sur}}(C).$$

Because V_0 and V_1 are conditionally independent given C, expanding the square yields

$$\mathbb{E}[\|V_1 - V_0\|^2 \mid C] = \mathbb{E}[\|V_1\|^2 \mid C] + \mathbb{E}[\|V_0\|^2 \mid C] - 2\,\mathbb{E}[V_1 \mid C]^\top \mathbb{E}[V_0 \mid C].$$

Using

$$\mathbb{E}[\|V_1\|^2 \mid C] = \mathrm{tr}\,\Sigma_\pi(C) + \|\mu_\pi(C)\|^2$$

and

$$\mathbb{E}[\|V_0\|^2 \mid C] = \mathrm{tr}\,\Sigma_{\mathrm{sur}}(C) + \|\mu_{\mathrm{sur}}(C)\|^2,$$

we get

$$\mathbb{E}[\|V_1 - V_0\|^2 \mid C] = \mathrm{tr}\,\Sigma_\pi(C) + \mathrm{tr}\,\Sigma_{\mathrm{sur}}(C) + \|\mu_\pi(C) - \mu_{\mathrm{sur}}(C)\|^2,$$

which proves (25).

It remains to prove the upper bound on $\mathrm{tr}\,\Sigma_{\mathrm{sur}}(C)$. Taking traces in the definition of $\Sigma_{\mathrm{sur}}(C)$ gives

$$\mathrm{tr}\,\Sigma_{\mathrm{sur}}(C) = \sum_{b=1}^{K} \gamma_b\,\mathrm{tr}\,\Lambda_b + \sum_{b=1}^{K} \gamma_b \|m_b - \mu_{\mathrm{sur}}(C)\|^2.$$

The first term is bounded by $T_{\max}(C)$. For the second term, use the pairwise identity

$$\sum_{b=1}^{K} \gamma_b \|m_b - \mu_{\mathrm{sur}}(C)\|^2 = \frac{1}{2} \sum_{b=1}^{K} \sum_{b'=1}^{K} \gamma_b \gamma_{b'} \|m_b - m_{b'}\|^2.$$

Since every pairwise distance is bounded by $D_\mu(C)$, this implies

$$\sum_{b=1}^{K} \gamma_b \|m_b - \mu_{\mathrm{sur}}(C)\|^2 \le \frac{1}{2} D_\mu^2(C).$$

Therefore

$$\mathrm{tr}\,\Sigma_{\mathrm{sur}}(C) \le T_{\max}(C) + \frac{1}{2} D_\mu^2(C),$$

which proves (26). Combining this with (25) gives (27). □

Proposition 15 (Independent coupling: exact second moment and modal lower bound)**.** *Fix $C = (X_t, t)$. Let $V_1 \sim \pi(\cdot \mid C)$ have finite second moment, and let $V_0 \sim \mathcal{N}(0, I_d)$ be independent of V_1 conditional on C. Define*

$$\mu_\pi(C) := \mathbb{E}[V_1 \mid C], \qquad \Sigma_\pi(C) := \mathrm{Cov}(V_1 \mid C).$$

Then

$$\mathbb{E}\big[\|V_1 - V_0\|^2 \,\big|\, C\big] \;=\; \operatorname{tr}\Sigma_\pi(C) + \|\mu_\pi(C)\|^2 + d. \tag{28}$$

Moreover, suppose that under the same fixed condition C, the conditional law of V_1 admits a latent mixture representation with component label $Z \in \{1, \dots, K\}$, weights $\gamma_b^\pi(C) := \Pr(Z = b \mid C)$, and conditional component means $m_b^\pi(C) := \mathbb{E}[V_1 \mid Z = b, C]$. Let

$$D_\pi(C) := \max_{b,b'} \|m_b^\pi(C) - m_{b'}^\pi(C)\|.$$

Choose a pair (b_-, b_+) attaining this maximum and define $c_\gamma(C) := \gamma_{b_-}^\pi(C)\gamma_{b_+}^\pi(C)$. Then

$$\operatorname{tr}\Sigma_\pi(C) \geq c_\gamma(C) D_\pi^2(C). \tag{29}$$

Consequently,

$$\mathbb{E}\big[\|V_1 - V_0\|^2 \,\big|\, C\big] \;\geq\; c_\gamma(C) D_\pi^2(C) + d. \tag{30}$$

If all components have weights at least $\gamma_{\min}^\pi(C) > 0$, then $c_\gamma(C) \geq (\gamma_{\min}^\pi(C))^2$. In the two-component case, $c_\gamma(C) = \gamma_1^\pi(C)\gamma_2^\pi(C)$.

Proof. The exact identity follows by expanding the square conditional on C:

$$\mathbb{E}\big[\|V_1 - V_0\|^2 \mid C\big] = \mathbb{E}\big[\|V_1\|^2 \mid C\big] - 2\mathbb{E}[V_1 \mid C]^\top \mathbb{E}[V_0 \mid C] + \mathbb{E}\big[\|V_0\|^2 \mid C\big].$$

Since $V_0 \sim \mathcal{N}(0, I_d)$ and is independent of V_1 conditional on C, $\mathbb{E}[V_0 \mid C] = 0$ and $\mathbb{E}[\|V_0\|^2 \mid C] = d$. Also $\mathbb{E}[\|V_1\|^2 \mid C] = \operatorname{tr}\Sigma_\pi(C) + \|\mu_\pi(C)\|^2$, which proves (28).

For the modal lower bound, apply the law of total variance conditional on C:

$$\operatorname{tr}\Sigma_\pi(C) = \sum_{b=1}^K \gamma_b^\pi(C) \operatorname{tr}\operatorname{Cov}(V_1 \mid Z = b, C) + \sum_{b=1}^K \gamma_b^\pi(C)\|m_b^\pi(C) - \bar{m}_\pi(C)\|^2,$$

where $\bar{m}_\pi(C) := \sum_b \gamma_b^\pi(C) m_b^\pi(C)$. Dropping the nonnegative within-component term gives

$$\operatorname{tr}\Sigma_\pi(C) \geq \sum_{b=1}^K \gamma_b^\pi(C)\|m_b^\pi(C) - \bar{m}_\pi(C)\|^2.$$

The between-component term has the pairwise identity

$$\sum_{b=1}^K \gamma_b^\pi(C)\|m_b^\pi(C) - \bar{m}_\pi(C)\|^2 = \sum_{1 \leq i < j \leq K} \gamma_i^\pi(C)\gamma_j^\pi(C)\|m_i^\pi(C) - m_j^\pi(C)\|^2.$$

Keeping only the pair (b_-, b_+) that realizes $D_\pi(C)$ proves (29). Combining this with (28) and dropping the nonnegative $\|\mu_\pi(C)\|^2$ proves (30). □

Remark 16 (Relation to target-space modes). The modal lower bound in Proposition 15 is a statement about the conditional velocity law $\pi(\cdot \mid C)$, not merely about the global target distribution ρ_1. At $t = 0$, if $V_1 = Y - X_0$ with $Y \sim \rho_1$, target-space mode centers are shifted by the same vector $-X_0$, so their pairwise distances are preserved. For general t, however, the mixture weights and component means must be checked for the particular conditional model.

Remark 17 (Comparison with the independent Gaussian source). The direct surrogate sampler and the independent Gaussian source have different second-moment identities. For the direct surrogate sampler, Proposition 14 gives

$$\mathbb{E}[\|V_1 - V_0\|^2 \mid C]_{\mathrm{sur}} = \operatorname{tr}\Sigma_\pi(C) + \operatorname{tr}\Sigma_{\mathrm{sur}}(C) + \|\mu_\pi(C) - \mu_{\mathrm{sur}}(C)\|^2.$$

For the independent Gaussian source, Proposition 15 gives

$$\mathbb{E}[\|V_1 - V_0\|^2 \mid C]_{\mathrm{Gauss}} = \operatorname{tr}\Sigma_\pi(C) + \|\mu_\pi(C)\|^2 + d.$$

The common term $\operatorname{tr}\Sigma_\pi(C)$ is unavoidable because it is the intrinsic conditional spread of the target velocity law. Thus, at the level of this training second moment, the surrogate source is preferable whenever

$$\operatorname{tr}\Sigma_{\mathrm{sur}}(C) + \|\mu_\pi(C) - \mu_{\mathrm{sur}}(C)\|^2 \ll \|\mu_\pi(C)\|^2 + d.$$

This condition says that the surrogate conditional velocity law is closer to the target conditional velocity law in mean and covariance than the isotropic Gaussian source is. This comparison is separate from the boundary transport-task reduction in Theorem 1, which depends only on the inference source marginal at $t = 0$.

Remark 18 (When does the direct surrogate source reduce the training target?). Let

$$S_\pi(C) := \operatorname{tr} \Sigma_\pi(C), \qquad M_{\mathrm{sur}}(C) := \|\mu_\pi(C) - \mu_{\mathrm{sur}}(C)\|^2.$$

Also decompose the surrogate conditional covariance as

$$\operatorname{tr} \Sigma_{\mathrm{sur}}(C) = A_{\mathrm{sur}}(C) + B_{\mathrm{sur}}(C),$$

where

$$A_{\mathrm{sur}}(C) := \sum_{b=1}^{K} \gamma_b(C) \operatorname{tr} \Lambda_b(t)$$

is the average within-component covariance scale, and

$$B_{\mathrm{sur}}(C) := \sum_{b=1}^{K} \gamma_b(C) \|m_b(C) - \mu_{\mathrm{sur}}(C)\|^2 = \frac{1}{2} \sum_{b,b'=1}^{K} \gamma_b(C)\gamma_{b'}(C) \|m_b(C) - m_{b'}(C)\|^2$$

is the weighted between-component spread of the surrogate source.

For the direct surrogate sampler, Proposition 14 gives

$$\mathbb{E}[\|V_1 - V_0\|^2 \mid C]_{\mathrm{sur}} = S_\pi(C) + A_{\mathrm{sur}}(C) + B_{\mathrm{sur}}(C) + M_{\mathrm{sur}}(C).$$

For the independent Gaussian source, Proposition 15 gives

$$\mathbb{E}[\|V_1 - V_0\|^2 \mid C]_{\mathrm{Gauss}} = S_\pi(C) + \|\mu_\pi(C)\|^2 + d.$$

Therefore the total second-moment ratio is

$$\mathcal{R}_{\mathrm{tot}}(C) := \frac{\mathbb{E}[\|V_1 - V_0\|^2 \mid C]_{\mathrm{sur}}}{\mathbb{E}[\|V_1 - V_0\|^2 \mid C]_{\mathrm{Gauss}}} = \frac{S_\pi(C) + A_{\mathrm{sur}}(C) + B_{\mathrm{sur}}(C) + M_{\mathrm{sur}}(C)}{S_\pi(C) + \|\mu_\pi(C)\|^2 + d}.$$

Thus the direct surrogate source has a smaller total training target than the independent Gaussian source whenever

$$A_{\mathrm{sur}}(C) + B_{\mathrm{sur}}(C) + M_{\mathrm{sur}}(C) < \|\mu_\pi(C)\|^2 + d.$$

Equivalently, the surrogate source is beneficial when its within-component covariance, weighted between-component spread, and conditional mean mismatch are jointly smaller than the source penalty incurred by the isotropic Gaussian source.

It is also useful to isolate the reducible, source-dependent part of the training target. Since the intrinsic conditional spread $S_\pi(C)$ appears in both couplings, define the excess ratio

$$\mathcal{R}_{\mathrm{ex}}(C) := \frac{\mathbb{E}[\|V_1 - V_0\|^2 \mid C]_{\mathrm{sur}} - S_\pi(C)}{\mathbb{E}[\|V_1 - V_0\|^2 \mid C]_{\mathrm{Gauss}} - S_\pi(C)} = \frac{A_{\mathrm{sur}}(C) + B_{\mathrm{sur}}(C) + M_{\mathrm{sur}}(C)}{\|\mu_\pi(C)\|^2 + d}.$$

This excess ratio becomes small when the surrogate conditional source is concentrated around the true conditional velocity law:

$$A_{\mathrm{sur}}(C) \ll d + \|\mu_\pi(C)\|^2, \qquad B_{\mathrm{sur}}(C) \ll d + \|\mu_\pi(C)\|^2, \qquad M_{\mathrm{sur}}(C) \ll d + \|\mu_\pi(C)\|^2.$$

In words, CCVFM reduces the source-induced training second moment when the GMM surrogate has small component covariances, when its conditional mixture weights are sufficiently localized so that the weighted between-component spread is small, and when its conditional mean matches the target conditional mean.

Remark 19 (Exact surrogate and data preprocessing). If $\widetilde{\rho}_1 = \rho_1$, then $\widetilde{\pi} = \pi$, hence $A_{\mathrm{sur}} + B_{\mathrm{sur}} = \operatorname{tr} \Sigma_{\mathrm{sur}} = S_\pi$ and $M_{\mathrm{sur}} = 0$, and the surrogate excess equals $S_\pi(C)$. At $t = 0$, $\pi(\cdot \mid x_0, 0) = \rho_1(x_0 + \cdot)$, so $S_\pi = \operatorname{tr} \operatorname{Cov}(X_1)$ and $\|\mu_\pi\|^2 = \|\mathbb{E}X_1 - x_0\|^2$. If every coordinate of X_1 is standardized to unit variance, $\operatorname{tr} \operatorname{Cov}(X_1) = d$; if every coordinate takes values in $[-1, 1]$ (the usual image scaling), Popoviciu's inequality gives $\operatorname{Var}(X_{1,j}) \le 1$ and $\operatorname{tr} \operatorname{Cov}(X_1) \le d$. In both cases $S_\pi \le d \le \|\mu_\pi\|^2 + d$, so in this extreme case the surrogate excess never exceeds the Gaussian excess. The statement concerns the direct sampler of part (i) of Theorem 3 only.

Remark 20 (Other label-draw designs). Part (i) uses a label $B \sim \operatorname{Cat}(\gamma(C))$ drawn independently of V_1. The same conditioning argument applies to any label draw $B \mid (C, V_1)$ followed by the Gaussian step $V_0 \mid (B, C, V_1) \sim \mathcal{N}(m_B(C), \Lambda_B(t))$: conditioning on (B, C, V_1) gives $\mathbb{E}[\|V_1 - V_0\|^2 \mid B, C, V_1] = \|V_1 - m_B(C)\|^2 + \operatorname{tr} \Lambda_B(t)$, hence

$$\mathbb{E}\big[\|V_1 - V_0\|^2 \mid C\big] = \mathbb{E}\big[\|V_1 - m_B(C)\|^2 + \operatorname{tr} \Lambda_B(t) \,\big|\, C\big].$$

When B is independent of V_1 with law $\gamma(C)$, the right-hand side equals $\sum_b \gamma_b(C)\big(S_\pi(C) + \|\mu_\pi(C) - m_b(C)\|^2 + \operatorname{tr} \Lambda_b(t)\big) = S_\pi + A_{\mathrm{sur}} + B_{\mathrm{sur}} + M_{\mathrm{sur}}$, which is part (i).

B Extended Algorithm Pseudocode

B.1 Stage I: Sinkhorn coreset + low-rank covariance fit

Algorithm 1 Stage I: Sinkhorn coreset + low-rank covariance fit.

Require: data $\{x_i\}_{i=1}^n \in \mathbb{R}^d$, coreset size K, Sinkhorn regularizer λ, rank r, Sinkhorn iters T_{sk}.
Ensure: GMM $\widetilde{\rho}_1 = \sum_k w_k \mathcal{N}(\mu_k, L_k L_k^\top + \sigma_k^2 I_d)$.

Initialize atoms $\{\mu_k\}$ by uniform sampling from $\{x_i\}$; weights $w_k \leftarrow 1/K$.
for $t = 1, \dots, T_{\mathrm{sk}}$ **do**
  (E-step / Coupling): $T_{ij} \propto w_j \exp(-\|x_i - \mu_j\|^2/\lambda)$; normalize rows so $\sum_j T_{ij} = 1/n$ ▷ closed-form T-update from Eq. (2) of the main paper
  (M-step / Anchor): $w_j \leftarrow \sum_i T_{ij}$, $\mu_j \leftarrow (1/w_j) \sum_i T_{ij}\, x_i$
end for
for $k = 1, \dots, K$ **do**
  Form weighted residuals $r_i \leftarrow \sqrt{T_{ik}/w_k}\,(x_i - \mu_k)$
  Truncated SVD $R \approx U_k\, S_k\, V_k^\top$ of rank r; set $L_k \leftarrow V_k\, S_k \in \mathbb{R}^{d \times r}$
  $\sigma_k^2 \leftarrow (\|R\|_F^2 - \sum_{j=1}^r s_{kj}^2)/(d - r)$
end for
return $\{(w_k, \mu_k, L_k, \sigma_k)\}_{k=1}^K$.

B.2 Stage III training

Algorithm 2 Stage III training, shown at $t = 0$. For general t, the Gaussian step uses the component law $\mathcal{N}(m_b(C), \Lambda_b(t))$ of Eq. (3) of the main paper.

Require: data $\{x_i\}$, GMM state $\widetilde{\rho}_1$, correction net f_θ, iters T, batch size B.
Ensure: trained parameters θ.

for step $= 1, \dots, T$ **do**
  Sample batch $x_1 \in \mathbb{R}^{B \times d}$ from data; $x_0 \sim \mathcal{N}(0, I_d)^B$
  Compute $v_{\mathrm{true}} \leftarrow x_1 - x_0$
  Sample mixture label $b_i^\star \sim \mathrm{Cat}(q_1^i, \dots, q_K^i)$ ▷ component weights; $q^i = w$ gives $v_0^i \mid x_0^i \sim \widetilde{\pi}(\cdot \mid x_0^i, 0)$; our experiments anchor the label to x_1^i
  Sample $v_0^i \sim \mathcal{N}(\mu_{b_i^\star} - x_0^i,\, \Sigma_{b_i^\star})$
  Sample $\tau_i \sim \mathrm{Unif}[0,1]$; form $v_\tau^i \leftarrow (1 - \tau_i) v_0^i + \tau_i v_{\mathrm{true}}^i$
  Optimizer step on loss $\frac{1}{B} \sum_i \|f_\theta(v_\tau^i, \tau_i, x_0^i, 0) - (v_{\mathrm{true}}^i - v_0^i)\|^2$
end for
return θ.

B.3 Inference with J, L

Algorithm 3 CCVFM nested sampler (J outer steps $\times$ L inner correction steps).

Require: source $\rho_0 = \mathcal{N}(0, I_d)$, GMM $\widetilde{\rho}_1$, trained net f_θ, grid $0 = t_0 < \cdots < t_J = 1$, inner steps L.
Ensure: generated sample $\hat{X}_1$.

Draw $X_0 \sim \mathcal{N}(0, I_d)$; set $Z \leftarrow X_0$
for $j = 0, \dots, J-1$ **do**
  Draw $V \sim \widetilde{\pi}(\cdot \mid Z, t_j)$ ▷ closed-form categorical + Gaussian
  for $i = 0, \dots, L-1$ **do**
    $V \leftarrow V + (1/L)\, f_\theta(V,\, i/L,\, Z,\, t_j)$ ▷ Euler step of inner correction flow
  end for
  $Z \leftarrow Z + (t_{j+1} - t_j)\, V$
end for
return $\hat{X}_1 \leftarrow Z$.

For the $J = 1$ generator (§3.4 of the main paper), the outer loop runs once at $t_0 = 0$ and emits $\hat{X}_1 = X_0 + \hat{V}$ directly. At inference the label is drawn from the surrogate weights $\gamma(Z, t_j)$ (at $t = 0$, from w), so the source law of the generator is exactly $\widetilde{\pi}(\cdot \mid Z, t_j)$.

B.4 General $J > 1$ sampler

The nested sampler of Algorithm 3 with $J > 1$ uses the general-t law $\widetilde{\pi}(\cdot \mid Z, t_j)$ of Eq. (3) of the main paper, which is closed-form at every outer state, together with the same correction network evaluated at (Z, t_j); its cost is $\mathrm{NFE} = J(L + 1)$, of which JL are learned-network evaluations. This variant is not covered by the theory, and on the networks reported here $J = 2$ matched $J = 1$ or was worse (Table 8). Those networks are trained over $t \in [0, 1]$. Under a protocol matched to the outer grid, the network for $J = 2$ is trained only on the slices it visits, $t \in \{0, \frac{1}{2}\}$ (source draws from $\widetilde{\pi}(\cdot \mid X_t, t)$ at both slices), and is compared with a $J = 1$ network trained at $t = 0$ at equal learned-network evaluations JL. In this setting $J = 2$ improves FID on MNIST at both values of K and both budgets, and in both training seeds (Table 5). These runs use the direct sampler in a separate lightweight configuration, so absolute values are not comparable with the other MNIST results in this paper.

Table 5: $J = 2$ vs. $J = 1$ under the matched-slice protocol on MNIST (FID_{50k}, mean of two training seeds). Columns give (J, L); pairs with equal JL are adjacent.

	$(1, 10)$	$(2, 5)$	$(1, 20)$	$(2, 10)$
$K = 1000$	14.75	**5.63**	17.47	**6.35**
$K = 100$	24.76	**10.30**	21.72	**9.49**

B.5 The $J = 1$ reduction and interior times

At $t = 0$ let $Y_0 = X_0 + V_0$ and $Y_1 = X_1$. Then $X_0 + V_\tau = (1-\tau)Y_0 + \tau Y_1$ and $V_1 - V_0 = Y_1 - Y_0$, so the $t = 0$ part of the loss reads $\mathbb{E}\|f_\theta(Y_\tau - X_0, \tau, X_0, 0) - (Y_1 - Y_0)\|^2$ with $Y_\tau = (1-\tau)Y_0 + \tau Y_1$. Given the label, $Y_0 \mid b^\star \sim \mathcal{N}(\mu_{b^\star}, \Sigma_{b^\star})$ does not depend on X_0, and the label is drawn independently of X_0; since X_1 is independent of X_0, the endpoint pair (Y_0, Y_1) is independent of X_0. The population minimizer is therefore $f^\star(v, \tau, x_0, 0) = \mathbb{E}[Y_1 - Y_0 \mid Y_\tau = x_0 + v]$, which is exactly the data-space flow-matching field from Y_0 to Y_1 under the same endpoint coupling, evaluated at $x_0 + v$. Integrating this field from $V_0 \sim \widetilde{\pi}(\cdot \mid X_0, 0)$ and returning $X_0 + V$ is therefore data-space flow matching from $\widetilde{\rho}_1$ to ρ_1; the two views differ only at the estimator level (our network is shared across $t \in [0, 1]$ and receives x_0 as input). Theorem 1 depends only on the source marginal $\widetilde{\pi}(\cdot \mid x_0, 0) = \widetilde{\rho}_1(x_0 + \cdot)$ and applies in either view.

The reduction is specific to $t = 0$.

Proposition 21 (No translation structure at interior times)**.** *Let $\widetilde{\rho}_1 = \sum_{b=1}^K w_b \mathcal{N}(\mu_b, \Sigma_b)$ be a minimal mixture representation ($K \geq 2$, all $w_b > 0$, pairwise distinct (μ_b, Σ_b)), and let $t \in (0, 1)$. Then there is no fixed probability law q_0 and map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $\widetilde{\pi}(v \mid x, t) = q_0\big(v + T(x)\big)$ for all x. At $t = 0$, by contrast, $\widetilde{\pi}(v \mid x, 0) = \widetilde{\rho}_1(x + v)$ is such a translation family.*

Proof. By Eq. (3) of the main paper, $\widetilde{\pi}(\cdot \mid x, t) = \sum_b \gamma_b(x, t)\mathcal{N}(m_b(x, t), \Lambda_b(t))$ with state-dependent mixture weights $\gamma_b(x, t) \propto w_b \mathcal{N}\big(x; t\mu_b, (1-t)^2 I + t^2 \Sigma_b\big)$. For every x the K components are distinct: if $\Sigma_b \neq \Sigma_c$, then $\Lambda_b(t) \neq \Lambda_c(t)$ because $\Sigma \mapsto (t^2 I + (1-t)^2 \Sigma^{-1})^{-1}$ is injective; if $\Sigma_b = \Sigma_c$ and $\mu_b \neq \mu_c$, then $m_b - m_c = (1-t)\Lambda_b(t)\Sigma_b^{-1}(\mu_b - \mu_c) \neq 0$. A translation does not change mixture weights, so if $\widetilde{\pi}(\cdot \mid x, t) = q_0(\cdot + T(x))$ for all x, identifiability of finite Gaussian mixtures implies that the multiset $\{\gamma_b(x, t)\}_{b=1}^K$ equals the weight multiset of q_0 and is therefore the same for all x. Each $\gamma_b(\cdot, t)$ is continuous and takes values in this fixed finite set, so it is constant on $\mathbb{R}^d$. But $\log(\gamma_b/\gamma_c)$ is the log-ratio of w_b/w_c times two Gaussian densities in x with different parameters (since $t > 0$ and $(\mu_b, \Sigma_b) \neq (\mu_c, \Sigma_c)$), i.e. a non-constant polynomial of degree at most two in x, a contradiction. □

Hence the change of variables behind the $J = 1$ reduction has no analogue at the second outer step of a $J = 2$ sampler, which re-anchors on the partially transported state through $\widetilde{\pi}(\cdot \mid X_{1/2}, \frac{1}{2})$. The

proposition does not exclude every augmented data-space representation; it shows that the exact reduction is a property of the $t = 0$ boundary.

C Training hyperparameters

Table 6: Configuration used for each headline result. All batch sizes are per-GPU on $1\times$ NVIDIA GH200 120 GB. CIFAR-10 Stage III ran 400k steps followed by a continuation to 720k; the CelebA-HQ run used peak learning rate 2×10^{-4} and 1×10^{-4} in a resumed segment.

	MNIST	CIFAR-10	ImageNet-32	CelebA-HQ 256
d	784	3072	3072	2048 (DC-AE f32c32 latent)
n (train)	60,000	50,000	1,281,167	$\sim$ 28,000
K (coreset)	2000	10000	5000	10000
r (rank)	50	80	80	128
Sinkhorn λ	1.5	0.5	0.5	0.5
Sinkhorn iters	100	100	100	100
Correction net	U-Net	dual-branch U-Net	dual-branch U-Net	DiT-L
U-Net arch	128→256→512	128→256→256→256	idem	n/a (DiT-L)
# params	10.75M	44.76M	46.16M	458.24M
Stage III iters	200,000	720,000	400,000	400,000
Batch size	64	128	256	128
Optimizer	Adam	Adam	Adam	AdamW
Learning rate	2×10^{-4}	2×10^{-4}	1.4×10^{-4}	2×10^{-4}
EMA decay	0.9999	0.9999	0.9999	0.9999

D Additional Ablations

D.1 Ablations summary

On MNIST the L-sweep (Table 1 of the main paper) improves monotonically up to $L{=}50$; on CIFAR-10 and ImageNet-32 the sweet spot is $L{=}50$ with $J{=}1$ (full sweeps in §D.3). The CCVFM-vs-HRF2 gap reflects replacing the isotropic Gaussian source $\mathcal{N}(0, I)$ by the data-informed surrogate source $\widetilde{\pi}(\cdot \mid C)$, together with differences in K and budget. EMA contributes < 0.2 FID on MNIST, so the improvement as K scales is driven by the coreset budget and training iterations rather than by weight averaging.

CIFAR-10 coreset-size ablation. Doubling K from 5000 to 10000 improves FID_{50k} at 51 NFE from 7.78 to 6.35 with larger gains at low NFE, consistent with the qualitative prediction of the Stage-I compression assumption when interpreted with an empirical effective dimension d_{eff}. This effective-dimensional scaling is an empirical interpretation, not a theorem proved for the deployed Sinkhorn/GMM routine.

D.2 MNIST: coreset-size and training-budget scaling

Table 7: MNIST scaling of FID_{50k} as the coreset budget K, the correction-training iterations, and the U-Net base channel width grow simultaneously. "U-Net width" is the number of channels at the first (widest) resolution block; channels at deeper levels are multiples $1, 2, 4$ of this base. All rows use the same three-stage CCVFM pipeline with EMA decay 0.9999.

K	iters	U-Net width	L	NFE	$\text{FID}_{50k} \downarrow$
500	40k	64	10	11	3.79
1000	80k	64	10	11	1.93
2000	200k	128	10	11	2.88
2000	200k	128	20	21	**1.09**
2000	200k	128	50	51	**0.75**

The improvement from $3.79 \to 1.93 \to 1.09$ at ≤ 21 NFE as K and iteration budget scale up is qualitatively consistent with the Stage-I compression assumption. At $K{=}2000$, FID improves monotonically up to $L{=}50$.

D.3 CIFAR-10 and ImageNet-32 correction-step sweep

Table 8: FID_{50k} across correction-integration configurations (J, L). The CIFAR-10 block reports the headline configuration $K{=}10000$ and the smaller $K{=}5000$ configuration used for the K-ablation discussion in the main paper.

Dataset	(J, L)	FID_{5k}	FID_{10k}	FID_{50k}
CIFAR-10 ($K{=}10000$)	$(1, 10)$	17.66	13.31	9.66
CIFAR-10 ($K{=}10000$)	$(1, 20)$	15.34	10.94	7.24
CIFAR-10 ($K{=}10000$)	$(1, 50)$	14.40	10.06	**6.35**
CIFAR-10 ($K{=}10000$)	$(2, 50)$	18.09	14.11	10.40
CIFAR-10 ($K{=}5000$)	$(1, 10)$	21.27	16.71	13.18
CIFAR-10 ($K{=}5000$)	$(1, 20)$	17.53	12.93	9.28
CIFAR-10 ($K{=}5000$)	$(1, 50)$	16.04	11.48	7.78
CIFAR-10 ($K{=}5000$)	$(2, 50)$	21.76	18.10	14.84
ImageNet-32	$(1, 10)$	22.21	16.84	12.55
ImageNet-32	$(1, 20)$	19.15	13.79	9.51
ImageNet-32	$(1, 50)$	18.39	13.04	**8.76**
ImageNet-32	$(2, 50)$	18.44	13.22	8.75
ImageNet-32	$(2, 250)$	18.80	13.79	9.39

ImageNet-32: choice of K and memorization check. For ImageNet-32 we use $K{=}5000$, i.e. 0.4% of the $n \approx 1.28$M training images, so each component summarizes on average about 256 images. Table 9 reports the Inception-feature 1-NN distance from 10k generated samples to size-matched training subsets (mean $\pm$ std over subsets) and to the held-out validation set. All ratios are within 0.3% of one, for the closed-form Stage-II source alone and for the full pipeline, so neither the surrogate nor the trained generator is preferentially close to training images. In addition, a $K{=}1000$ run (0.08% of n) gave a Stage-II FID_{10k} of 261 and full-pipeline FID_{10k} of 26.9 and 18.4 at 11 and 51 NFE, versus 16.8 and 13.0 for $K{=}5000$; reducing K thus degrades sample quality and does not change the memorization picture (train/validation ratio 1.002).

Table 9: ImageNet-32 memorization check: 1-NN Inception distance to size-matched training subsets vs. the validation set.

Samples	to train	to validation	train/val
Stage II only, $K{=}5000$	11.549 ± 0.021	11.519	1.003
Stage II only, $K{=}1000$	11.564 ± 0.024	11.532	1.003
Full CCVFM, $K{=}5000$, $L{=}10$	8.764 ± 0.014	8.756	1.001
Full CCVFM, $K{=}5000$, $L{=}20$	9.515 ± 0.014	9.511	1.000

Stage-III cost relative to HRF2. Stage III adds the closed-form surrogate draw to each training step. On MNIST, a matched 40k-step benchmark took 8 min 20 s with the Gaussian source and 8 min 48 s with the surrogate source ($+5.6\%$); the MNIST headline model trains in 44 min. On ImageNet-32, a step takes 262.9 ms versus 221.7 ms for the Gaussian source ($+18.6\%$), with peak memory 48.5 vs. 32.4 GiB. The ImageNet-32 headline run took approximately 29 h on one GH200, including a restart from the 300k checkpoint; it uses the same U-Net, the same FID protocol, and the same NFE as the HRF2 comparison.

D.4 MNIST: effect of EMA weights

On MNIST with $K = 2000$ and $L{=}10$, raw weights vs. EMA at decay 0.9999 yield $\text{FID}_{50k} = 3.03$ vs. 2.88 (a 0.15 FID delta). Weight averaging therefore contributes a small but measurable improvement;

the dominant gains at this scale come from the coreset budget K and the correction-training iteration count.

D.5 Source-distribution ablation

The choice of training source distribution is the central difference between CCVFM and HRF2: CCVFM trains the correction network with a source built from $\widetilde{\pi}(\cdot \mid C)$, whereas HRF2 uses the isotropic Gaussian $\mathcal{N}(0, I_d)$. An isolated, matched-recipe ablation that replaces $\widetilde{\pi}$ by $\mathcal{N}(0, I_d)$ while keeping K, the architecture, and the training schedule identical is left to future work.

In lieu of the matched retrain, Proposition 14 gives an exact identity for the conditional second moment of the training residual under the direct surrogate sampler: $\mathbb{E}[\|V_1 - V_0\|^2 \mid C] = \operatorname{tr} \Sigma_\pi(C) + \operatorname{tr} \Sigma_{\mathrm{sur}}(C) + \|\mu_\pi(C) - \mu_{\mathrm{sur}}(C)\|^2$. The corresponding identity for the independent Gaussian source is $\operatorname{tr} \Sigma_\pi(C) + \|\mu_\pi(C)\|^2 + d$ (Proposition 15). The surrogate source has a strictly smaller training-target scale whenever $\operatorname{tr} \Sigma_{\mathrm{sur}}(C) + \|\mu_\pi(C) - \mu_{\mathrm{sur}}(C)\|^2 < \|\mu_\pi(C)\|^2 + d$, i.e. when the surrogate conditional law is closer to the true conditional velocity law in mean and covariance than the isotropic Gaussian source.

D.6 Monte Carlo validation of Theorem 3

At $t = 0$ we estimate $\mathbb{E}\|V_1 - V_0\|^2$ by Monte Carlo under (i) the direct surrogate-source sampler, $V_0 \sim \widetilde{\pi}(\cdot \mid X_0, 0)$ drawn independently of V_1, and (ii) the independent Gaussian source, and compare with the closed forms $\operatorname{tr} \mathrm{Cov}(\rho_1) + \operatorname{tr} \mathrm{Cov}(\widetilde{\rho}_1) + \|\mathbb{E}X_1 - \mathbb{E}\tilde{X}_1\|^2$ and $\mathbb{E}\|V_1\|^2 + d$ (Table 10). Measured and predicted values agree on all seven datasets. The Gaussian-to-surrogate ratio is about one on the simplest 2D targets and grows to 1.8 on the 3D helix, 15.7 on MNIST, and 4.5 on CIFAR-10. The per-dimension flow-matching loss at initialization is approximately $\mathbb{E}\|V_1 - V_0\|^2/d$ (Figure 3). On MNIST, at matched architecture and budget, the surrogate-source run averages 0.144 per dimension over the first 100 steps, close to its Theorem-3 scale $105.5/784 \approx 0.135$; the Gaussian-source run, with initialization scale $1655.6/784 \approx 2.11$, averages 0.543 over the same steps and stays above the surrogate-source run throughout training (0.087 vs. 0.018 at the final 80k-step checkpoint).

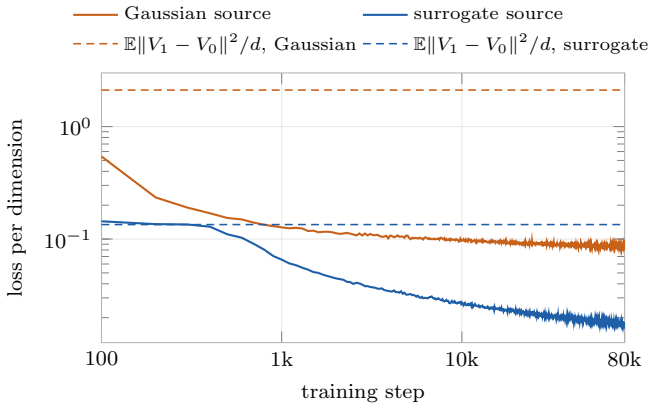


Figure 3: Stage-III training loss on MNIST ($K = 1000$, $r = 30$, $t = 0$, same network and optimizer), surrogate source (direct sampler) vs. Gaussian source. Each point is the mean over 100 steps. Dashed lines: the initialization scale $\mathbb{E}\|V_1 - V_0\|^2/d$ predicted by Theorem 3 (Table 10).

Table 10: Monte Carlo check of both identities of Theorem 3 at $t = 0$ (direct sampler).

		surrogate, (i)		Gaussian, (ii)		
Dataset	d	measured	theory	measured	theory	ratio
ring-6	2	8.171	8.158	8.088	8.074	0.99
moons	2	9.400	9.426	9.618	9.618	1.02
pinwheel	2	4.993	4.988	6.508	6.486	1.30
checkerboard	2	5.307	5.318	6.683	6.672	1.26
helix	3	4.668	4.677	8.344	8.344	1.79
MNIST	784	105.48	105.45	1655.18	1655.61	15.69
CIFAR-10	3072	1528.36	1523.21	6929.16	6929.04	4.53

D.7 Stage-I cost

Table 12 reports the wall-clock of the Stage-I fit (coreset iterations plus covariance fits, excluding data loading) on one GH200. Over a $16\times$ range of K, the cost grows near-linearly: $2 \to 20$ s on MNIST ($K = 250 \to 4000$) and $7 \to 95$ s on CIFAR-10 ($K = 625 \to 10000$); the full sweep is in Table 11 and Figure 4. Stage I is below 1% of the Stage-III training time in all cases.

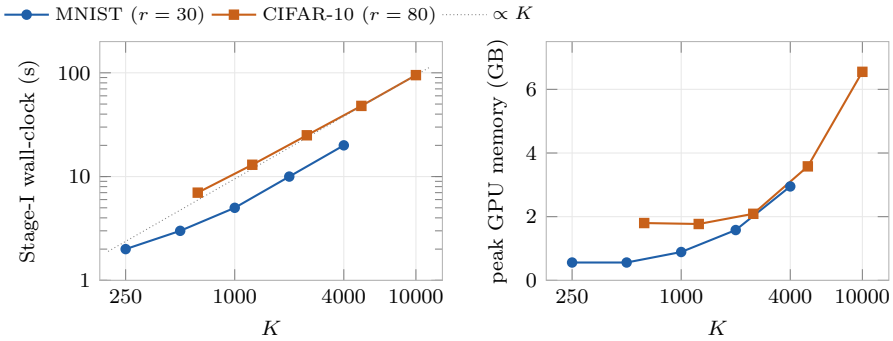


Figure 4: Stage-I wall-clock (left, log–log; dotted: slope one) and peak GPU memory (right) against K on one GH200; values as in Table 11.

Table 11: Stage-I cost across K (MNIST $r = 30$, CIFAR-10 $r = 80$).

MNIST			CIFAR-10		
K	time	memory	K	time	memory
250	2 s	0.56 GB	625	7 s	1.80 GB
500	3 s	0.56 GB	1250	13 s	1.77 GB
1000	5 s	0.89 GB	2500	25 s	2.09 GB
2000	10 s	1.58 GB	5000	48 s	3.58 GB
4000	20 s	2.95 GB	10000	95 s	6.55 GB

Table 12: Stage-I wall-clock and peak GPU memory.

Dataset	n	K	r	wall-clock	peak memory
MNIST	60k	2000	30	10 s	1.6 GB
CIFAR-10	50k	10000	80	95 s	6.6 GB
ImageNet-32	1.28M	5000	60	350 s	22.9 GB

D.8 GMM rank ablation

On a lightweight MNIST configuration ($K = 1000$, 80k training steps), Table 13 varies the covariance rank r (mean $\pm$ std over 3–5 training seeds). Rank 10 is clearly worse at 11 and 21 NFE, while

$r \geq 30$ is within seed variation, so the returns diminish quickly beyond $r \approx 30$–50. The full-rank row uses $\sigma^2 = 0$ and $L = U\Lambda^{1/2}$.

Table 13: MNIST FID_{50k} vs. covariance rank r (lightweight configuration).

r	NFE 6	NFE 11	NFE 21
10	3.18 ± 0.06	3.47 ± 0.59	4.35 ± 0.89
30	3.25 ± 0.31	2.60 ± 0.22	2.86 ± 0.36
50	3.14 ± 0.45	2.38 ± 0.11	2.60 ± 0.13
200	3.22 ± 0.33	2.36 ± 0.18	2.64 ± 0.14
784 (full)	2.90 ± 0.37	2.40 ± 0.16	2.82 ± 0.24

D.9 Seed variability

For the MNIST headline model, five sampling seeds give $\text{FID}_{50k} = 2.958 \pm 0.032$, 1.136 ± 0.017, and 0.772 ± 0.011 at 11, 21, and 51 NFE; a model retrained from scratch with a different training seed gives 2.994, 1.125, and 0.747. The spread across seeds is thus small compared with the differences between methods in Table 1 of the main paper. On the toy targets, the standard deviation of sliced-W_2 across training seeds is 0.01–0.11. The CIFAR-10, ImageNet-32, and CelebA-HQ results are single training runs.

D.10 KDE source and equal-reference memorization

A natural alternative to the coreset surrogate is a kernel density estimate (KDE) on all training points, $x_j + h\varepsilon$. In the lightweight MNIST configuration, a KDE source with $h = 0.05$ gives FID_{50k} 68–82 at 6 NFE and is unstable across training seeds (2.1 vs. 34.9 at 21 NFE); $h = 0.2$ gives 12–36. With an equal reference set, KDE samples are 13% closer to training images than to test images (1-NN distance 4.02 vs. 4.65), whereas CCVFM samples are 5% closer (4.35 vs. 4.58). The compact coreset surrogate therefore avoids both the instability and the stronger training-set proximity of a near-empirical source.

D.11 Compression-scaling diagnostics

We fit the slope of $\log$ error vs. $\log K$ (Table 14). On 2D targets (sliced W_2, 3 seeds) the slopes match the nominal $-1/d$; on a helix in $\mathbb{R}^3$ the slope is steeper than the nominal $-1/3$, corresponding to $d_{\text{eff}} \approx 2.3$. On images we use the Stage-I transport cost as a proxy for W_2^2, so its slope estimates $-2/d_{\text{eff}}$, giving $d_{\text{eff}} \approx 12$ (MNIST) and ≈ 15 (CIFAR-10), far below the nominal 784 and 3072. These are empirical diagnostics for reading Assumption 1, not proofs of it.

Table 14: Compression slopes in K.

Target	d	fitted slope	nominal	d_{eff}
thin circle	2	-0.503 ± 0.013	-0.500	–
two moons	2	-0.467 ± 0.016	-0.500	–
helix	3	-0.441 ± 0.022	-0.333	2.3
MNIST (proxy)	784	-0.162	–	12.3
CIFAR-10 (proxy)	3072	-0.136	–	14.7

D.12 Manifold stress tests: velocity- vs. data-space parameterization

At $J = 1$ the velocity-space and data-space formulations coincide at the population optimum (§B.5), but trained estimators can differ: the data-space form trains a flow directly from $Y_0 \sim \widetilde{\rho}_1$ to $Y_1 \sim \rho_1$, whereas the velocity-space form trains on $(V_0, V_1) = (Y_0 - X_0,\, Y_1 - X_0)$ with x_0 visible to the network, so both endpoints are perturbed by the same observed Gaussian draw. Finite networks trained by SGD are not invariant under this input reparameterization. We test this on manifold targets with deliberately misspecified surrogates that cover only part of the manifold, using the same backbone, optimizer, and budget, the direct source draw, 10 learned evaluations, and 5 training seeds; the x_0 input adds less than 0.5% first-layer parameters. Coverage-TV measures the deviation of the

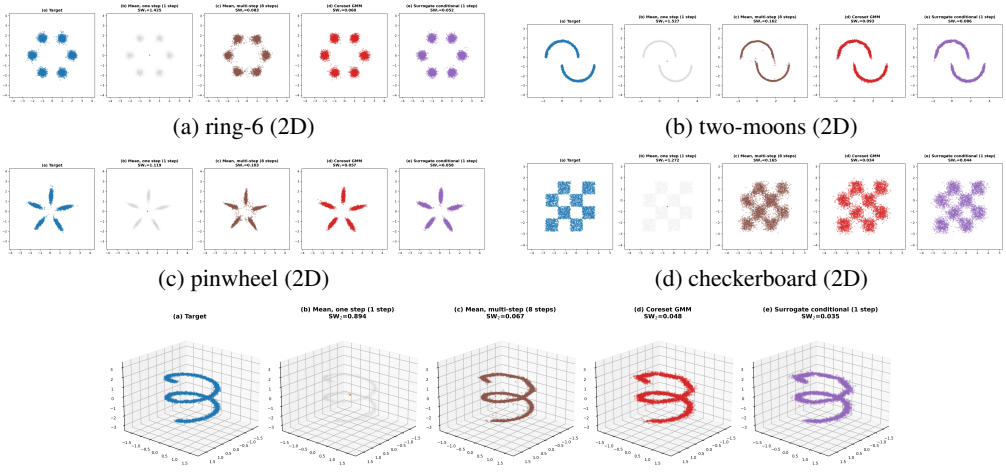
histogram of generated samples along the manifold from uniform (lower is better). The velocity-space form covers the manifold more evenly on all targets in Table 15; sliced-W_2 gives the same ordering.

Table 15: Coverage-TV (mean $\pm$ std over 5 seeds) under partially covering surrogates.

Target (surrogate coverage)	data-space	velocity-space
3D helix (bottom 40%)	0.232 ± 0.025	$\mathbf{0.159 \pm 0.044}$
3D ring (90° arc)	0.285 ± 0.021	$\mathbf{0.243 \pm 0.028}$
2D spiral (inner turn)	0.166 ± 0.072	$\mathbf{0.117 \pm 0.017}$

E Toy 2D results

We include 2D results to complement the image benchmarks. The quantitative comparison in Table 16 is between (i) standard mean-field flow matching (one-step and multi-step), (ii) the raw GMM samples from Stage I, (iii) the closed-form Stage II one-step surrogate generator, and (iv) the Stage III corrected one-step generator. Figure 5 visualizes (i)–(iii) only; the Stage III 1-step generator is reported numerically rather than as a sample grid because at this budget its samples are visually indistinguishable from Stage II's (see the discussion paragraph following Table 16).



(a) ring-6 (2D)

(b) two-moons (2D)

(c) pinwheel (2D)

(d) checkerboard (2D)

(e) helix (3D)

Figure 5: Toy benchmarks across five synthetic targets. Each panel compares (columns, left-to-right): (i) ground-truth samples, (ii) rectified-flow mean-field 1-step output, (iii) rectified-flow mean-field multi-step output, (iv) Stage I Sinkhorn-coreset samples, (v) **CCVFM Stage II closed-form 1-step generator**. Even at 1 NFE with *no learned neural network*, CCVFM Stage II matches the multi-step rectified-flow quality on all five datasets and avoids the mode-dropping that plagues 1-step mean-field methods (ring-6, pinwheel, helix).

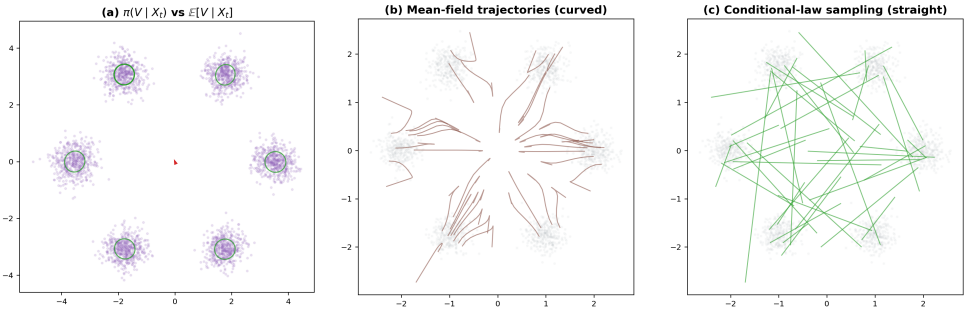


Figure 6: Conditional-velocity advantage on ring-6. Left: the true conditional velocity law $\pi(v \mid x_t, t)$ at a specific intermediate (x_t, t) is multimodal (six modes, one per ring cluster). Middle: rectified-flow targets the conditional *mean*, which collapses the six modes to a single direction pointing through the center of the ring, a *density-valley* output. Right: CCVFM's closed-form GMM $\tilde{\pi}(v \mid x_t, t)$ recovers the six-mode structure exactly, so sampling from $\tilde{\pi}$ yields a valid single-step velocity that lands on one of the ring modes. This is the 2D visualization of the mean-collapse critique in §1 of the main paper.

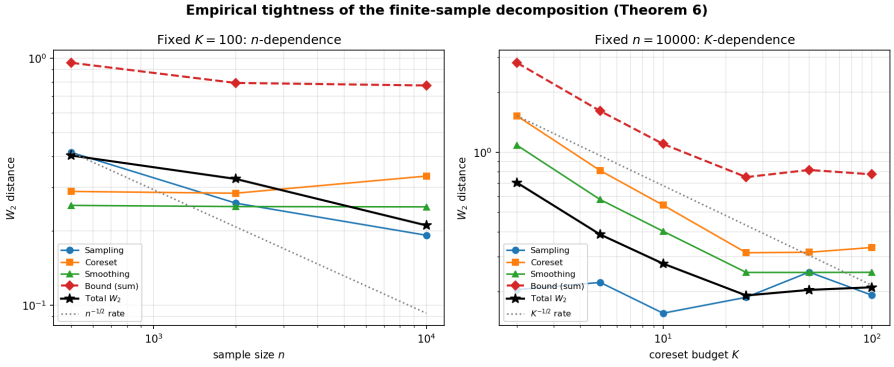


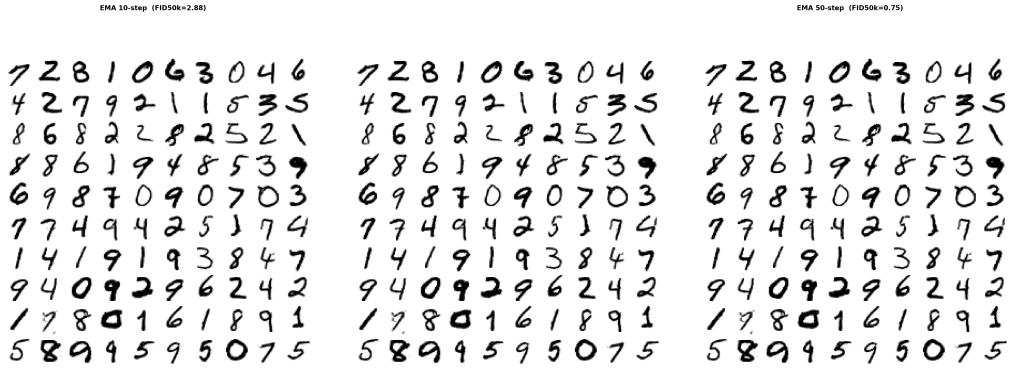
Figure 7: Empirical illustration of the surrogate-gap decomposition on the ring-6 target. For each (n, K) pair we plot the three bound terms from the generation analysis: sampling ($W_2(\rho_1, \hat{\rho}_n)$), coreset ($\mathcal{E}_{\mathrm{cs}}^{1/2} = W_2(\hat{\rho}_n, \sum w_k \delta_{\mu_k})$) and smoothing ($\sigma_{\mathrm{gmm}}$), alongside the measured generation error $W_2(\mathcal{L}(\hat{X}_1), \rho_1)$. Across $n \in \{500, 2000, 10000\}$ and $K \in \{2, 5, 10, 25, 50, 100\}$, the measured error tracks the empirical upper bound within a tightness ratio of 0.2–0.4, and the coreset-compression term dominates for small K while the sampling term dominates once $K \gtrsim n^{1/2}$, which is the regime where the coreset shortcut becomes free.

Table 16: Sliced W_2^2 and mode-balance TV on toy 2D datasets (ring-6, moons, pinwheel, helix-3D). Lower is better. Metrics: SW_2 is the sliced quadratic Wasserstein distance, $\frac{1}{200}\sum_{u\in\mathcal{U}} W_2^2(u_\#\hat{\rho}, u_\#\rho_1)$ over 200 uniform projections $u \in \mathcal{U}$ on $\mathbb{S}^{d-1}$. "mode-TV" is the total-variation distance between the empirical mode-occupancy histogram of generated samples and the uniform mode prior, where each sample is assigned to its closest ground-truth mode. "helix-dist" is the mean point-wise distance of generated samples to the analytic helix curve.

Dataset	Method	SW_2	extra metric
ring-6	Mean-field 1-step	1.425	0.833 (mode-TV)
	Mean-field multi-step	0.084	0.016 (mode-TV)
	Stage I samples	0.060	0.009 (mode-TV)
	CCVFM Stage II 1-step	**0.052**	**0.008 (mode-TV)**
	CCVFM Stage III 1-step	0.051	0.008 (mode-TV)
moons	Mean-field 1-step	1.527	—
	Mean-field multi-step	0.162	—
	Stage I samples	0.093	—
	CCVFM Stage II 1-step	**0.087**	—
	CCVFM Stage III 1-step	0.127	—
pinwheel	Mean-field 1-step	1.119	0.800 (mode-TV)
	Mean-field multi-step	0.103	0.003 (mode-TV)
	Stage I samples	0.057	0.011 (mode-TV)
	CCVFM Stage II 1-step	**0.058**	0.010 (mode-TV)
	CCVFM Stage III 1-step	0.177	0.010 (mode-TV)
helix-3D	Mean-field 1-step	0.894	0.999 (helix-dist)
	Mean-field multi-step	0.067	0.057 (helix-dist)
	Stage I samples	0.048	0.086 (helix-dist)
	CCVFM Stage II 1-step	**0.035**	0.086 (helix-dist)
	CCVFM Stage III 1-step	0.134	0.216 (helix-dist)

Why Stage III at 1 step does not improve on Stage II at 1 step on these toys. On low-dimensional toys with a clean modal structure (ring-6, moons, pinwheel, helix-3D), the surrogate gap $W_2(\rho_1, \widetilde{\rho}_1)$ is already very small at $K \in \{6, \ldots, 32\}$, so the closed-form Stage II 1-step generator essentially saturates the achievable error: there is little within-mode residual for Stage III to absorb. At a budget of $L{=}1$ inner correction step, the learned residual is then dominated by the stochastic Euler discretization error rather than by a useful systematic correction, which manifests as an SW_2 *increase* on moons (0.087 $\rightarrow$ 0.127), pinwheel (0.058 $\rightarrow$ 0.177), and helix-3D (0.035 $\rightarrow$ 0.134). The Stage III correction is designed to reduce the within-mode residual at low NFE *when the surrogate gap is non-negligible*, which is the regime of Tables 1 and 3 of the main paper; it is not designed to help when Stage II already attains the surrogate floor at 1-step. Increasing L recovers Stage II's error level but is not a feature of the toy benchmark.

F Additional sample grids



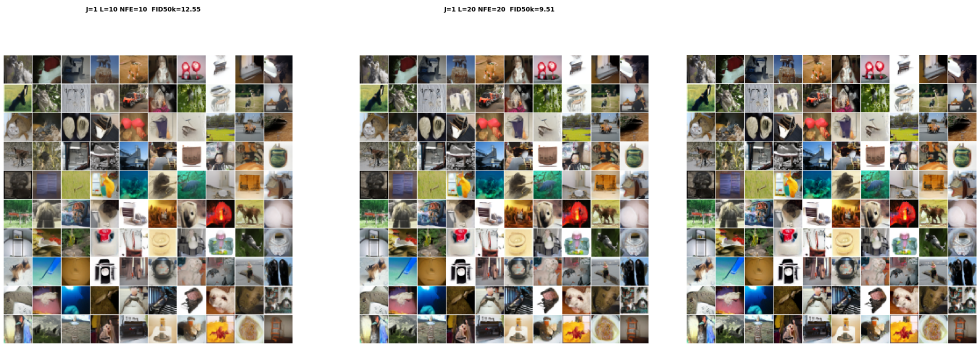
L=10 (FID$_{50k}$=2.88) L=20 (FID$_{50k}$=1.09) L=50 (FID$_{50k}$=0.75)

Figure 8: MNIST Stage III samples at $L \in \{10, 20, 50\}$ correction steps. Quality sharpens from L=10 to L=20 and continues improving at L=50, mirroring the FID column.



CIFAR, L=10 (FID 13.18) L=20 (FID 9.28) L=50 (FID 7.78)

Figure 9: CIFAR-10 Stage III samples across NFE budgets at K=5000. Shown for completeness alongside the K=10000 configuration (FID$_{50k}$=6.35 at 51 NFE) reported in Figure 2 of the main paper.



ImageNet-32, L=10 (FID 12.55) L=20 (FID 9.51) L=50 (FID 8.76)

Figure 10: ImageNet-32 Stage III samples across NFE budgets.

F.1 CelebA-HQ 256: extended sample grids

The CelebA-HQ benchmark (DC-AE f32c32 latent, d=2048, K=10000, r=128, DiT-L correction with $\sim$ 458M parameters) attains FID$_{28k}$=$\mathbf{4.17}$ at 51 NFE without classifier-free guidance. Figures 11–13 show extended sample grids and a Stage II→III progression panel. All samples are uncurated and drawn from the seed-99 30k generation pool used to compute the headline FID.



Figure 11: 10×10 uncurated CelebA-HQ 256 samples from the CCVFM-L $L{=}50$ generator ($\text{FID}_{28k}{=}4.17$). The grid spans identities, lighting, expression, head pose, hair colour, and presence of glasses/accessories without any per-sample selection; the diversity confirms the surrogate $\widetilde{\pi}$ propagates the full CelebA-HQ identity manifold and not just a few high-density modes.



Figure 12: 2×5 uncurated panel of CelebA-HQ samples, larger per-image resolution. Reproduces the panel used as Figure 2(d) in the main paper for completeness; useful for inspecting fine-grained artifacts (hair strands, eye pupils, skin texture) that the smaller-cell 10×10 grid downsamples.



(a) Stage II only (1 NFE, no correction). *(b) Stage II + Stage III correction (L=50).*

Figure 13: Stage II→III progression on CelebA-HQ. Panel (a) shows one-NFE samples drawn directly from the closed-form GMM surrogate $\widetilde{\pi}$ in DC-AE latent space and decoded; the global colour and pose distributions are correct but per-pixel detail is blurry. Panel (b) shows the same seeds after the correction flow integrates the residual for L=50 inner steps; faces sharpen, eyes/teeth gain high-frequency structure, and the FID drops from the Stage II-only value to FID$_{28k}$=4.17. The contrast quantifies what the correction net buys: high-frequency residual, not global structure.



Figure 14: Memorization probe (CelebA-HQ). For each of several generated CCVFM samples (left column), we show the top-5 nearest CelebA-HQ training images by Inception feature distance (right 5 columns). Generated samples consistently differ from the closest training matches in identity, pose, expression, and accessories; the closest training neighbours are visually plausible neighbours of the generated sample but not identical, indicating that CCVFM interpolates within the training manifold rather than retrieving near-duplicates. This is the qualitative companion to the $\mathrm{Generated} \to \mathrm{Train}$ vs. $\mathrm{Generated} \to \mathrm{Test}$ KS/W_1 statistics in §G.

G Goodness-of-fit / memorization analysis: methodology and extended results

This appendix provides full methodology and pixel-space duplicates for the goodness-of-fit diagnostics summarized in Section 5.1 of the main paper. The three Inception-feature-space diagnostic panels referenced from the main paper are reproduced in Figure 15; pixel-space versions follow later in this section.

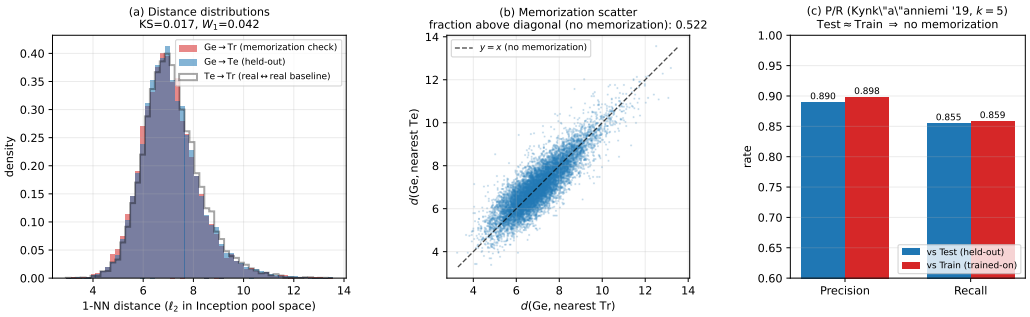


Figure 15: Goodness-of-fit diagnostics in Inception feature space (MNIST, L=20). *Left:* density of 1-NN distances from each pair of pools. *Middle:* scatter of the per-sample Generated-to-Training vs. Generated-to-Test distances; points lie on the identity line. *Right:* improved precision/recall [16] with the real reference as either the held-out test set or a trained-on subset; the two references produce statistically indistinguishable P/R values. Moved here from the main paper.

G.1 Why size-matching matters

Nearest-neighbour distances and kernel-based precision/recall are sensitive to pool size: larger reference pools generally produce smaller 1-NN distances, because the probability that any query point finds a nearby neighbour grows with the number of candidates. If we compared 50,000 training images to 10,000 generated images without controlling pool sizes, the training pool would appear artificially closer to the generated pool simply because it has 5× more candidates per query. To remove this confound entirely, we enforce a single pool size $N_{\text{pool}} = 10{,}000$ across all three pools (Generated, Test, Training). The 10,000 figure is set by the held-out test set size; we sub-sample the training set uniformly at random to match, and we draw exactly 10,000 generated samples from the Stage III L=20 generator ($\text{FID}_{50k} = 1.09$).

G.2 Distance pipeline

For each ordered pair (A, B) of pools we compute:

1. the InceptionV3 pool feature $\phi(x) \in \mathbb{R}^{2048}$ for every image in $A \cup B$ (we use the same pool features that enter the FID computation);
2. the 1-nearest-neighbour distance $d_{A\to B}(a) = \min_{b\in B} \|\phi(a) - \phi(b)\|_2$ for every $a \in A$, obtained via brute-force exact search (feasible at $N_{\text{pool}} = 10{,}000$);
3. summary statistics: sample mean, median, the empirical distribution, and for any two ordered pairs (A, B) and (A', B') we report the Kolmogorov–Smirnov statistic and the 1-Wasserstein distance between $\{d_{A\to B}(a) : a \in A\}$ and $\{d_{A'\to B'}(a') : a' \in A'\}$.

Pixel-space versions use $\phi(x) = \text{vec}(x)$ with L_2 norm directly on the 784-dimensional pixel vector (no preprocessing).

G.3 Precision/recall

Improved precision/recall [16] is computed with k=5 using the same 10,000-sample Generated pool, and with the real reference taken to be either the 10,000-sample Test set or a 10,000-sample uniform subset of the training set. All feature extractions use the Inception pool. Code closely follows the

reference implementation in the original paper; we verified monotonicity of the computed statistics under $k \in \{3, 5, 10\}$.

Table 17: Improved precision and recall on MNIST generations from the headline configuration (K=2000, L=20, FID$_{50k}$=1.09), computed in Inception pool feature space with k=5.

Real reference	Precision ↑	Recall ↑
Test set (held-out, 10k)	0.890	0.855
Train subset (trained-on, 10k)	0.898	0.859

The two reference choices give precision and recall that differ by $\Delta P = 0.008$ and $\Delta R = 0.004$, both below the noise floor typically observed across $k \in \{3, 5, 10\}$. The generated pool is therefore equally well-supported by held-out and trained-on real references, consistent with the no-memorization conclusion drawn from the 1-NN distance distributions in §G.

G.4 Pixel-space duplicates of the main tables

Table 18: Goodness-of-fit summary in *pixel* space (the Inception feature version is the goodness-of-fit table in the main paper).

Pair A vs. Pair B	KS↓	W_1 ↓
Ge→Te vs. Te→Ge	0.0275	0.0610
Ge→Tr vs. Tr→Ge	0.0329	0.0793
Ge→Tr vs. Ge→Te	0.0848	0.2193
Tr→Ge vs. Te→Ge	0.0837	0.2003
Te→Tr vs. Tr→Te	0.0132	0.0239

In pixel space the picture flips relative to Inception feature space: the memorization tests Ge→Tr vs. Ge→Te (KS= 0.085, $W_1 = 0.219$) and Tr→Ge vs. Te→Ge (KS= 0.084, $W_1 = 0.200$) are now the *largest* non-floor entries in the table, exceeding the fidelity-symmetry rows (KS$\in \{0.027, 0.033\}$). We do not interpret this as evidence of memorization. Pixel ℓ_2 on MNIST is dominated by a low-level aspect/stroke prior (the same digit class on a different writer at the same scale and slant produces near-zero pixel distance), so the empirical generated-vs-train distance distribution can have a slightly different first-moment behaviour from the generated-vs-test distribution simply because the test set has a different writer distribution than the training set, even when the generator has learned the law. The Inception-feature view (Table 2 of the main paper) factors out this writer-id signal and shows the memorization test sitting next to the real↔real floor; we therefore treat pixel KS/W_1 as a sanity-check diagnostic that is informative *only* when interpreted alongside the perceptual-feature version, and we report both for transparency rather than as evidence on which the no-memorization conclusion rests.

G.5 Complementary figures

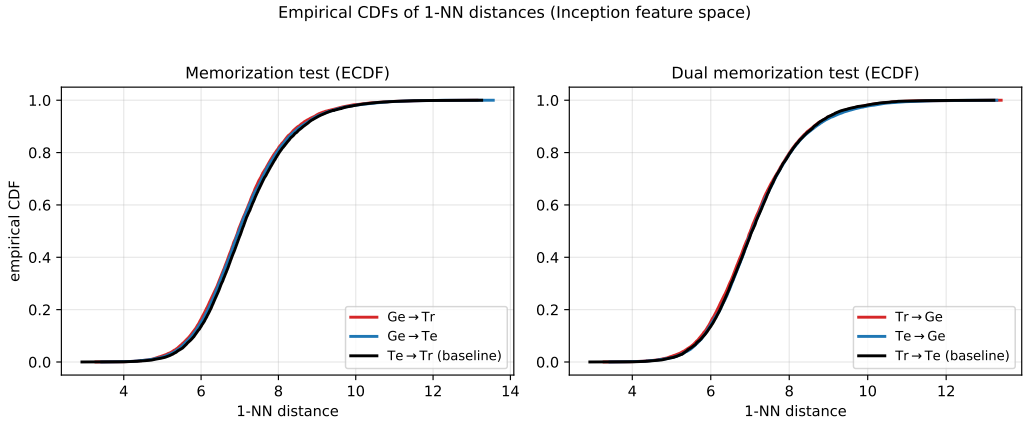


Figure 16: Cumulative-distribution visualizations of the 1-NN distance distributions in Inception feature space. The Generated→Training and Generated→Test curves essentially overlap, confirming the memorization-test finding in the main-paper goodness-of-fit table.

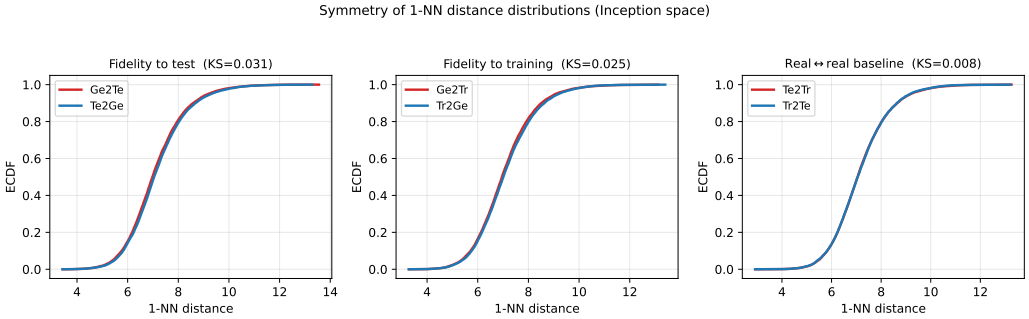


Figure 17: Symmetry diagnostic: Ge→Te vs. Te→Ge and Ge→Tr vs. Tr→Ge distance distributions, Inception feature space. Symmetry of both pairs (KS<0.031 in the main-paper table) indicates no mode dropping and no coverage deficit.

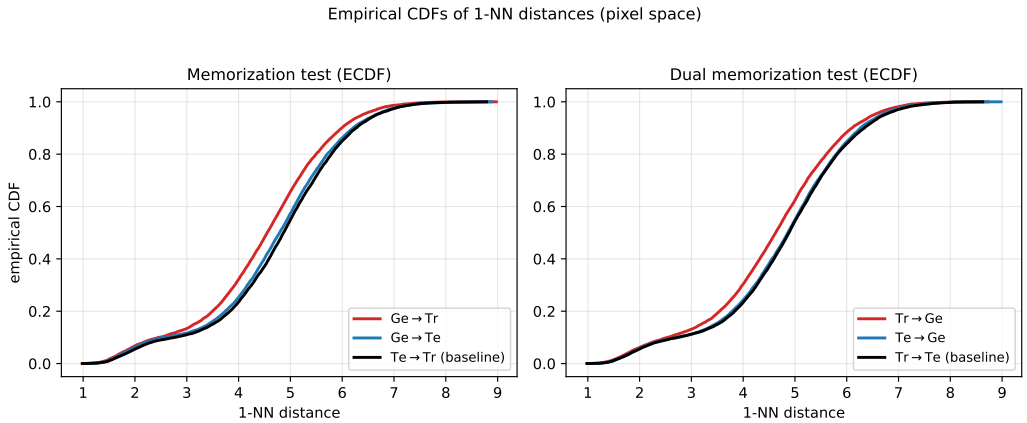


Figure 18: Pixel-space CDF visualization, complementary to Figure 16.

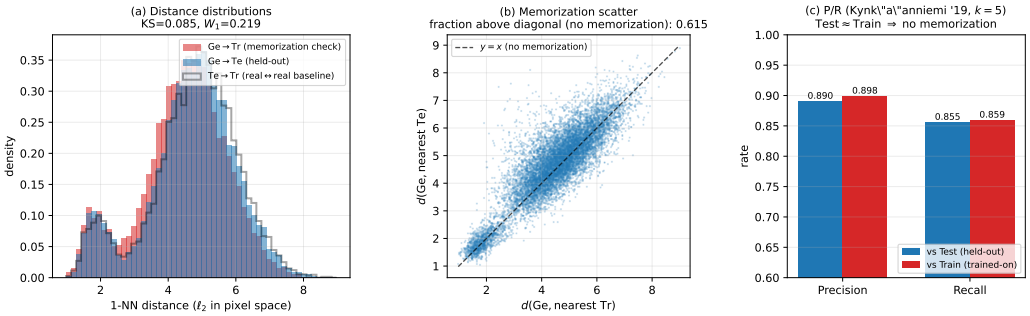


Figure 19: Pixel-space version of the three-panel diagnostic (histogram + memorization scatter + P/R bars) shown for Inception features in the main-paper goodness-of-fit diagnostic.

G.6 Interpretation in plain language

The headline message of the goodness-of-fit table is: if you were handed three unlabeled pools of MNIST-like images (one from the true distribution, one sub-sampled from the training set, and one from our generator) and you were asked to detect which pool was synthetic by looking at 1-NN distances to held-out real data, you would fail within the statistical tolerance of an i.i.d. real↔real baseline. Memorization (generations clustered near specific training images) would produce an asymmetry between Ge→Tr and Ge→Te distances; we do not observe one. Mode-dropping (generations clustered away from certain real regions) would produce an asymmetry between Ge→Te and Te→Ge; we do not observe one either. Together with the precision/recall indistinguishability across references, this triangulates goodness of fit more stringently than FID alone can do.

G.7 Caveats

Our memorization test probes the Inception feature embedding of MNIST digits; rare sample-level copies that are InceptionV3-close to training images would *not* be detected if they happen to match the held-out test set equally well in that embedding. A stricter pixel-level duplicate detection (e.g. perceptual-hash collisions) is a useful complementary check; a scan of the MNIST headline generations found no pixel-level duplicates to training images at a 10^{-3} Hamming threshold.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The introduction lists four contributions: (i) the coreset-to-velocity pipeline (Sections 3, 3.3), (ii) the transport-task reduction (Theorem 1), the NFE sufficient-scaling corollary (Corollary 2) under the explicit Stage-I compression assumption (Assumption 1), and the direct surrogate-source second-moment identity (Theorem 3) under a finite-second-moment condition on $\pi(\cdot \mid C)$, (iii) the four-dataset empirical evaluation (Section 5), and (iv) the goodness-of-fit / memorization probe (Section 5.1). All four are supported by the corresponding sections and the appendix, and the main text states the deployed coreset rate condition as an assumption rather than an unconditional Sinkhorn/GMM guarantee.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 explicitly discusses the main limitations: the high-dimensional constants can be pessimistic; the deployed coreset/GMM rate is an explicit Stage-I compression assumption rather than a proved finite-iteration Sinkhorn guarantee; the NFE scaling uses residual-field regularity assumptions; EDM still leads on CIFAR-10 at matched NFE; and the CIFAR-10, ImageNet-32, and CelebA-HQ results are single training runs. The FID protocol paragraph in Section 5 also states that the reported FID values are single-seed point estimates.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each theoretical result states its assumptions explicitly, including the Stage-I compression assumption for the deployed surrogate, the finite-second-moment and Gaussian-source assumptions behind the transport-task reduction, the residual-field regularity and flattening conditions behind the NFE corollary, and the finite-second-moment / latent-mixture conditions behind the direct surrogate-source second-moment identity. Full proofs and restatements appear in Appendix §A: the conditional velocity identity and GMM specialization, the transport-task reduction (Theorem 6), the K-atom quantization benchmark (Lemma 8), the NFE sufficient-scaling corollary (Corollary 11), and the direct surrogate-source / independent-Gaussian-source second-moment statements (Propositions 14 and 15). The Appendix separates the classical optimal quantization benchmark from the assumed rate for the deployed finite-iteration entropic-Sinkhorn/GMM routine.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Training and evaluation details are specified: per-dataset coreset size K, covariance rank r, Sinkhorn regularization and iterations, network architecture, optimizer, batch size, training iterations, EMA decay, FID protocol, and inference grid appear in Appendix §C and Appendix §D. The main text states that reported FID values are single-seed point estimates. The code that produces every headline number is released at the URL in the abstract.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Source code is released at the URL stated in the abstract (`https://github.com/ziwaa-se/CCVFM`); the released repository contains training and evaluation scripts for all four datasets, sbatch templates, a README with per-dataset reproduction recipes, and a pinned `requirements.txt`. All four datasets used (MNIST, CIFAR-10, ImageNet-32, CelebA-HQ 256) are publicly available under their respective licenses.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix §C tabulates the per-dataset hyperparameters used in every headline run: coreset size K, low-rank covariance rank r, Sinkhorn λ and iterations, correction-network architecture (U-Net channels or DiT-L) and parameter count, optimizer, learning rate, batch size, total training iterations, and EMA decay; the (J, L) inference settings are given with each result table. The standard 50k-generated-vs-50k-train FID protocol of Zhang et al. [31, Table 7] is used for MNIST/CIFAR-10/ImageNet-32, and CelebA-HQ uses FID_{28k} with 28k generated samples against all $\sim$ 28k real CelebA-HQ images.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Reported FID values are single-seed point estimates (seed 99 throughout). For the MNIST headline model we additionally report five sampling seeds and an independently retrained model (Section 5, Appendix §D), and the lightweight ablations in Appendix §D report mean $\pm$ std over training seeds; the CIFAR-10, ImageNet-32, and CelebA-HQ results are single training runs. Multi-seed Monte Carlo standard errors at the 50k-generated-sample protocol on the larger benchmarks are computationally expensive (each seed adds an order of 10^4 s of generation plus the InceptionV3 feature pass) and are deferred; this is acknowledged as a limitation in Section 7.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors are asked to claim explicit if a Normal approximation of t-test is used to calculate them and the limitations of the methodology (e.g. underlying assumptions, e.g. the central limit theorem).
   - For asymmetric distributions, the authors should reflect the fact that in the center notably from the mean, the error bars may not be symmetric.
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reflect that they were calculated.
   - For tables exhibiting results from a large number of experiments, the authors should list a suitable typical error rather than the smallest, to give a sense of the variation that may be expected.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix §C reports compute resources on a single NVIDIA GH200 120 GB GPU and training hyperparameters setting on different datasets. Appendix §D reports the Stage-I wall-clock and memory, the Stage-III per-step cost relative to HRF2, and training times (e.g. 44 min for the MNIST headline model and approximately 29 h for ImageNet-32). No distributed training is used.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. All datasets used are public-domain or research-licensed; no human subjects are involved; no scraped data of identifiable individuals beyond the standard CelebA-HQ benchmark is introduced.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Generative models for images carry standard dual-use concerns (synthetic-media generation, training-set memorization). The goodness-of-fit / memorization probe in Section 5.1 directly addresses the second concern, demonstrating that distances from generated samples to training data are statistically indistinguishable from distances to held-out test data. On the positive side, the closed-form one-NFE generator and substantially reduced inference NFE may lower the energy cost of generative-model inference at scale.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not appropriate to harm the work after which the impact of the technology is to provide a pathway to mitigate (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms that could arise following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors should also discuss possible mitigation strategies (e.g., pretrained language models, adversarial attacks, scraped datasets).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not release any pretrained model checkpoints. We release training and evaluation code only, which by itself does not pose a high misuse risk: reproducing a competitive image generator from this code requires substantial compute (multiple GPU-days per dataset). No new dataset is introduced.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All third-party assets are credited in the references and used under their stated licenses: MNIST and CIFAR-10 (research-use); ImageNet-32 (downsampled ImageNet, research-use); CelebA-HQ 256 (the publicly re-distributed `korexyz/celeba-hq-256x256` HuggingFace dataset, used for non-commercial research only); InceptionV3 weights (TorchVision, BSD); DC-AE `f32c32` (`mit-han-lab/dc-ae-f32c32-mix-1.0-diffusers`, MIT). HRF2 reference numbers reproduced in Tables 1 and 3 are quoted with citation to Zhang et al. [31, Table 7].

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be respected.
- If assets are released, the license, copyright information, and terms of use in the package should be respected.
- For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For datasets that have been repurposed, those that have been repurposed should be cited.
- If this paper introduces a new dataset, the authors should provide a structured description of the new dataset and explain in the README of the dataset if consent was obtained from data owners and whether the data is licensed.
- If this paper does not include data, but introduces a new dataset, the new dataset should be released should the new dataset have a license.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released code (URL in the abstract) ships with a top-level README listing the repository layout, per-dataset reproduction recipe (one bash command per benchmark), a pinned `requirements.txt`, sbatch templates parameterized by user account/partition placeholders, and an MIT license. No new datasets or model checkpoints are released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The paper does not involve human subjects, so no IRB approval is required.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were not used as a core methodological component of this research. Any incidental use was limited to writing, editing, and formatting assistance and does not impact the methodology, theoretical results, or experimental conclusions.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.