TOWARD GENERALIZATION FROM MEMORIZED DIFFUSION MODELS WITH HOUSEKEEPING-CURRENT INJECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a sampling protocol that reduces copying in memorized diffusion models while retaining the large-scale structure of the data distribution. We describe memorization with a relative free-energy model, in which narrow local minima around individual training examples lie within broader energy wells. Using the Entropic Bound (EB) together with the continuity equation, we distinguish density-changing transport from housekeeping currents that preserve a prescribed reference density. This motivates housekeeping-current injection (HCI), which adds such a current during sampling to steer trajectories away from memorized paths. In the energy picture, HCI helps trajectories leave the narrow wells around training examples; the learned landscape itself is not changed, and the diffusion model is not retrained. We also add a simple transport correction to alleviate the concentration that can occur with HCI alone. A 2D ring illustrates the landscape and HCI sampling; experiments on an 8D Gaussian mixture model, CelebA-32 and CelebA-64 evaluate trained models that show strong memorization. Compared with recent sampling-based methods, HCI substantially reduces the memorization fraction f_{mem} at a smaller loss of sample quality.

1 INTRODUCTION

Diffusion models generate samples using a learned reverse process (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021b). With finite training data, prolonged training can lead to copies of training examples (Gu et al., 2023; Baptista et al., 2025; Bonnaire et al., 2025). In a relative free-energy model, this copying is linked to sharp local peaks in the learned density around training examples (Pham et al., 2025; Jeon et al., 2025), while the surrounding landscape may still contain information about the broader structure of the data distribution.

Our idea is to use this information to guide sampling trajectories beyond individual training examples. Simply moving samples away from them is not enough: the added velocity should help samples escape the narrow wells and move within the broader ones, without blurring images or removing detail.

Entropy production measures the irreversibility of a diffusion process and the associated loss of information (Seifert, 2012), and it is a weighted norm of the probability current. The current can be decomposed into transport and housekeeping components (Dechant and Sasa, 2018; Ito, 2024): transport produces the required density evolution with the least entropy production, while housekeeping changes sample paths without changing that evolution. We use the Entropic Bound (EB), which relates current observables to entropy production, to derive two speed-limit-like inequalities in the transport and housekeeping directions. Based on them, we propose *housekeeping-current injection* (HCI), which adds a housekeeping current during reverse sampling to reduce memorization without significantly damaging sample quality.

Contributions. (i) A free-energy picture in which broad structure coexists with narrow training-point minima, supported by a trained 2D ring (SI E). (ii) An EB-based split of the probability current into transport and housekeeping, with a bound for each sector. (iii) Exact statements on reference preservation, KL neutrality, the response of the learned density and level-set motion. (iv) A training-

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

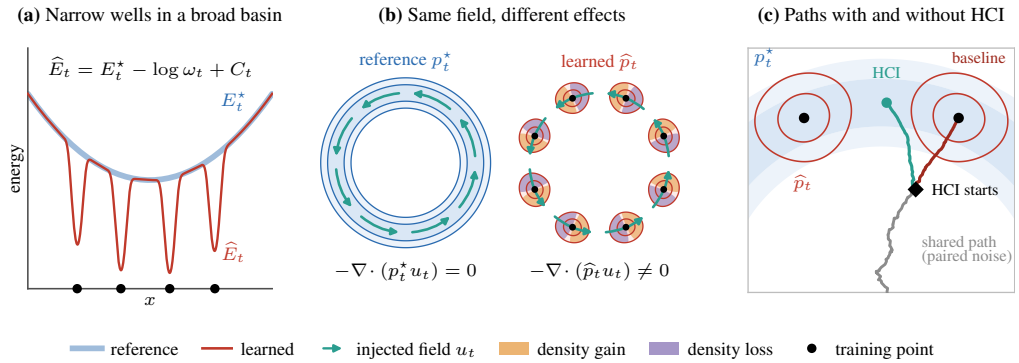


Figure 1: **Housekeeping-current injection (HCI) moves learned probability without changing the reference. (a)** Schematic 1D landscape: memorization adds narrow wells at the training points to the broad reference basin ($\omega_t = \widehat{p}_t/p_t^\star$). **(b)** On a ring, the same rotation field $u_t = K_2 x/\|x\|$ leaves $p_t^\star$ unchanged but shifts the peaks of $\widehat{p}_t$. **(c)** Close-up of two training points with schematic paired paths (same initial state and noise): once HCI starts, the baseline is captured by a training point, whereas the HCI path escapes along the ring.

free sampler built from a reference fitted to data, evaluated on frozen memorized backbones against churn, KTS and CPSample, including a replication on three CelebA configurations under a selection rule fixed in advance. Derivations and implementation details are in the supplementary information (SI).

Related sampling interventions. Churn adds stochasticity (Karras et al., 2022); KTS reshapes a flow-matching velocity (Li et al., 2026); higher-order Langevin methods use auxiliary variables (Sterling et al., 2026); CPSample guides samples away from training images with a classifier trained on random labels (Kazdan et al., 2025). HCI instead adds a new drift direction that preserves a reference. We test exact KTS on a native flow-matching ring (Lipman et al., 2023); for the GMM and CelebA score models we use a VP adaptation of KTS.

Notation. A star denotes the prescribed reference $p_t^\star$: ideally the VP-noised path of the data distribution $p^\star$, which is unknown, so the sampler of Section 4.4 replaces it by an estimate p_t^{ref} fitted to data. A hat denotes the uncontrolled learned marginal $\hat{p}_t$. The frozen network output $\hat{s}(x,t)$ need not equal $\nabla \log \hat{p}_t$; $s_t^\star = \nabla \log p_t^\star$. For drift b_t and state-independent diffusion covariance $2D_t$, the mean local velocity is $v_t = b_t - D_t \nabla \log p_t$ and $\partial_t p_t = -\nabla\cdot(p_t v_t)$, so $p_t v_t$ is a probability current, not the drift itself. In Sections 3 and 4, $t \in [0,T]$ runs in the direction of sampling; from Section 4.4 on we use the VP convention, in which $t = 0$ is the data end and sampling runs from $t = T$ to 0. Training progress is reported in update counts. SI A collects the notation.

2 Memorization in a free-energy landscape

At a fixed sampling time, write the reference and learned densities through dimensionless energies, $p_t^\star = Z_t^{-1} e^{-E_t^\star}$ and $\hat{p}_t = \hat{Z}_t^{-1} e^{-\hat{E}_t}$. Their density ratio $\omega_t = \hat{p}_t/p_t^\star$ relates the two energies,

$$\hat{E}_t(x) = E_t^\star(x) + \Delta E_t(x) + C_t, \qquad \Delta E_t(x) := -\log \omega_t(x), \tag{1}$$

with $C_t = \log Z_t - \log \hat{Z}_t$ constant in x. For a density q, the *relative free energy* is its Kullback–Leibler divergence to the reference,

$$\mathcal{F}_t[q] := D_{\mathrm{KL}}(q\|p_t^\star) = \int q E_t^\star \,\mathrm{d}x - \left(-\int q \log q \,\mathrm{d}x\right) + \log Z_t, \tag{2}$$

a mean energy minus an entropy. SI D.1 gives the algebra and the variational derivative used below.

If memorization makes ω_t large near a training example, ΔE_t is locally negative there and a sampling trajectory tends to end close to that example. We are interested in the case in which such narrow minima coexist with broad structure (Figure 1). HCI then changes the sampling paths without changing the trained network.

An energy landscape on a trained ring. We train an MLP score network on 16 equally spaced points on a unit circle. To examine the learned field, we integrate its radial and angular components to obtain the projected potentials E_{rad} and E_{ang}, defined in SI E. The mean angular component is removed before integration so that E_{ang} is periodic.

At noise level $\sigma = 0.05$, the radial potential has a minimum near radius one. As training proceeds from 10^4 to 10^5 updates, the copy fraction rises from 0.070 to 0.935, while the learned radius remains close to one. At the later checkpoint, the angular potential at the midpoint between neighboring training points is 3.76 higher than at the training points (Figure 6 and Table 1). The ring structure therefore remains in the learned field even after narrow angular wells have formed around individual training points. This motivates changing the motion along the ring rather than removing the radial restoring force.

3 From the entropic bound to transport and housekeeping currents

We use EB to bound the endpoint error and the score error by entropy production, and split the probability current into a density-changing transport part and a path-changing housekeeping part. We use this split as a guide: it shows which direction can change sampling paths without changing the reference density.

Let $p_t = p_t^\star$ be the exact reverse density on $\mathbb{R}^d$, $t \in [0, T]$, with probability velocity v_t, $\partial_t p_t + \nabla \cdot (p_t v_t) = 0$, and scalar diffusion coefficient $D(t) > 0$, $D_t = D(t)I$. For a probe field Λ_t, define the entropy production, the time-integrated current and the probe complexity,

$$\Sigma[v] := \int_0^T \mathbb{E}_{p_t}\left[\frac{\|v_t\|^2}{D(t)}\right] \mathrm{d}t, \quad J(v,\Lambda) := \int_0^T \mathbb{E}_{p_t}[\Lambda_t \cdot v_t]\,\mathrm{d}t, \quad C[\Lambda] := \int_0^T \mathbb{E}_{p_t}[D(t)\|\Lambda_t\|^2]\,\mathrm{d}t. \tag{3}$$

The Entropic Bound (Dechant and Sasa, 2018) relates the current observable to entropy production and probe complexity,

$$|J(v,\Lambda)|^2 \le \Sigma[v]\, C[\Lambda]. \tag{4}$$

Two choices of probe give two branches (proofs in SI B).

Endpoint branch. If p_t and $\hat{p}_t$ share the velocity v_t, the probe $\Lambda_t = r_t \nabla \psi_t$, with $r_t = (p_t - \hat{p}_t)/p_t$ and ψ_t a Kantorovich witness for $W_1(p_t, \hat{p}_t)$, gives $(\Delta W_1)^2/(TR_0) \le \int_0^T D(t)\,\mathbb{E}_{p_t}[\|v_t\|^2/D(t)]\,\mathrm{d}t$, where ΔW_1 is the change of $W_1(p_t, \hat{p}_t)$ over $[0,T]$ and R_0 the initial Pearson divergence. This recovers the speed–accuracy relation of Ikeda et al. (2025) (SI B.2).

Score-gap branch. Let $s_t = s_t^\star$ and $\hat{s}_t = \hat{\boldsymbol{s}}(\cdot, t)$. The probe $\Lambda_t = \hat{s}_t - s_t$ gives

$$\frac{(\Delta W_{\mathrm{bias}})^2}{T\mathcal{D}_{\mathrm{err}}} \le \Sigma[v], \quad \Delta W_{\mathrm{bias}} := \int_0^T \mathbb{E}_{p_t}[(\hat{s}_t - s_t)\cdot v_t]\,\mathrm{d}t, \quad \mathcal{D}_{\mathrm{err}} := \frac{1}{T}\int_0^T D(t)\mathbb{E}_{p_t}\|\hat{s}_t - s_t\|^2\,\mathrm{d}t. \tag{5}$$

It measures the score error accumulated along the reverse current rather than an endpoint distance; $T\mathcal{D}_{\mathrm{err}}$ is the score mismatch that appears in the likelihood relation of Kodama and Hinczewski (2025). If $\hat{s}_t = \nabla \log \hat{p}_t$, then $\hat{s}_t - s_t = -\nabla \Delta E_t$, which connects this branch to (1).

Transport and housekeeping currents. The weighted Helmholtz–Hodge decomposition (Ito, 2024) splits the velocity into a transport part and a housekeeping part, $v_t = v_t^{\mathrm{tr}} + v_t^{\mathrm{hk}}$ with $\nabla \cdot (p_t v_t^{\mathrm{hk}}) = 0$, which are orthogonal in $\langle a, b\rangle := \int_0^T \mathbb{E}_{p_t}[a_t \cdot b_t / D(t)]\,\mathrm{d}t$. Consequently,

$$\partial_t p_t = -\nabla \cdot (p_t v_t^{\mathrm{tr}}), \qquad \Sigma[v] = \Sigma[v^{\mathrm{tr}}] + \Sigma[v^{\mathrm{hk}}]. \tag{6}$$

Transport makes the density evolve with the least entropy production; housekeeping changes sample paths without changing the density. We write P^p_{tr} and P^p_{hk} for the two projections relative to a density p (SI B.4). The score gap splits accordingly, $\Delta W_{\mathrm{bias}} = \Delta W_{\parallel} + \Delta H_{\mathrm{hk}}$ with $\Delta W_{\parallel} := \int_0^T \mathbb{E}_{p_t}[(\hat{s}_t - s_t)\cdot v_t^{\mathrm{tr}}]\,\mathrm{d}t$ and $\Delta H_{\mathrm{hk}} := \int_0^T \mathbb{E}_{p_t}[(\hat{s}_t - s_t)\cdot v_t^{\mathrm{hk}}]\,\mathrm{d}t$, and EB yields one bound per sector,

$$\frac{(\Delta W_{\parallel})^2}{T\mathcal{D}_{\mathrm{err}}} \le \Sigma_{\mathrm{tr}}, \qquad \frac{(\Delta H_{\mathrm{hk}})^2}{T\mathcal{D}_{\mathrm{err}}} \le \Sigma_{\mathrm{hk}}. \tag{7}$$

The first measures the score error along density-changing transport, the second its component along the housekeeping current.

From the split to an intervention. A housekeeping current relative to the reference preserves the reference density, but it can change the memorized learned density: $\nabla\cdot(p_t^\star u_t)=0$ does not imply $\nabla\cdot(\hat{p}_t u_t)=0$. This suggests adding ηu_t during sampling to change the memorized sampling paths while preserving the reference. We call this housekeeping-current injection (HCI).

4 HOUSEKEEPING-CURRENT INJECTION

4.1 PROPERTIES OF HCI

Adding a current ηu_t to the reference mean velocity $v_t^\star = b_t^\star - D_t s_t^\star$ modifies the Fokker–Planck equation as

$$\partial_t p_t^\eta = -\nabla\cdot\left[(b_t^\star+\eta u_t)p_t^\eta - D_t\nabla p_t^\eta\right], \qquad p_0^\eta = p_0^\star, \tag{8}$$

where η sets the injection strength. Unlike stochastic interventions, HCI modifies the drift, not the diffusion.

Theorem 1 (Exact reference preservation)**.** *For a fixed strength $\eta\neq 0$, the density path is preserved if and only if the added probability current is divergence-free under the reference,*

$$p_t^\eta = p_t^\star \quad\Longleftrightarrow\quad \nabla\cdot(p_t^\star u_t)=0. \tag{9}$$

Proof sketch. Substituting $p_t^\eta = p_t^\star$ in (8) and subtracting the reference Fokker–Planck equation leaves $\eta\nabla\cdot(p_t^\star u_t)=0$; the converse follows from uniqueness (SI C).

Proposition 1 (KL neutrality)**.** *Let $\nabla\cdot(p_t^\star u_t)=0$ and let a density q evolve by $\partial_\tau q = -\eta\nabla\cdot(qu_t)$ at fixed t. Then*

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mathcal{F}_t[q]\Big|_{u_t} = \eta\int q\, u_t\cdot\nabla\log\frac{q}{p_t^\star}\,\mathrm{d}x = 0. \tag{10}$$

The proof integrates by parts after writing the integrand as $p_t^\star u_t\cdot\nabla(q/p_t^\star)$ (SI D). A housekeeping current therefore does not change the relative free energy of any density. It does, however, change a density that deviates from the reference:

Proposition 2 (The same current acts differently on the learned density)**.** *Let $\hat{p}_t = p_t^\star\omega_t$ and $\nabla\cdot(p_t^\star u_t)=0$. The injected drift contributes to the continuity equation of $\hat{p}_t$ the source*

$$-\nabla\cdot(\hat{p}_t u_t) = -p_t^\star u_t\cdot\nabla\omega_t, \tag{11}$$

which is nonzero unless ω_t is constant along u_t.

Conversely, a field that preserves the memorized marginal itself cannot change its population copy fraction (SI C); the useful freedom is preservation of a different reference.

4.2 EFFECT OF HCI ON A MEMORIZED MODEL

To use the free-energy picture, we freeze the reference at sampling time t and introduce an auxiliary time τ. The velocity that decreases $\mathcal{F}_t$ is $v_{t,\mathrm{eq}}^{\mathrm{tr}}[q] = -D_t\nabla(\delta\mathcal{F}_t/\delta q) = -D_t\nabla\log(q/p_t^\star)$. With a reference-preserving u_t, consider the relaxation, started from $q_0=\hat{p}_t$,

$$\partial_\tau q_\tau = \nabla\cdot\left[q_\tau D_t\nabla\log\frac{q_\tau}{p_t^\star}\right] - \eta\nabla\cdot(q_\tau u_t). \tag{12}$$

By Proposition 1, the injected current drops out of the free-energy balance,

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mathcal{F}_t[q_\tau] = -\int q_\tau\left(\nabla\log\frac{q_\tau}{p_t^\star}\right)^\top D_t\nabla\log\frac{q_\tau}{p_t^\star}\,\mathrm{d}x \le 0, \tag{13}$$

but it changes q_τ and hence the rate at which the transport term lowers the free energy. HCI makes no direct contribution to this decrease, but it can affect later relaxation by changing q_τ. This frozen-time model illustrates the roles of HCI and transport. The learned sampler need not follow the same relaxation.

4.3 CONSTRUCTED FIELDS AND TRANSPORT CORRECTION

We next construct a simple HCI field and explain why an additional transport correction can be useful in the learned sampler.

Constructed fields. Projecting a given field onto the housekeeping current requires solving a weighted Poisson problem in d dimensions at every t. We instead use *constructed fields* $u_t = -Ks_t^\star$ with a constant skew matrix K, which satisfy (9) by design and need no training given a closed-form reference score.

Proposition 3 (HCI moves samples within level sets)**.** *Let $u_t = -Ks_t^\star$ with a constant skew matrix K. Then $\nabla\cdot u_t = 0$ and $u_t \cdot s_t^\star = 0$, hence $\nabla\cdot(p_t^\star u_t) = 0$. Along the flow of u_t at fixed t, $p_t^\star(x)$ is constant, so the distribution of $\log p_t^\star(X)$ is unchanged for any law of X.*

Proof sketch. $\nabla\cdot u_t = -\sum_{ij} K_{ij}\partial_i s_j^\star = 0$ because $\nabla s_t^\star$ is a symmetric Hessian, and $u_t\cdot s_t^\star = -s_t^{\star\top} K s_t^\star = 0$ by skew symmetry. Then $\nabla\cdot(p_t^\star u_t) = p_t^\star(\nabla\cdot u_t + u_t\cdot s_t^\star) = 0$ and $\frac{\mathrm{d}}{\mathrm{d}s}p_t^\star(x(s)) = p_t^\star\, u_t\cdot s_t^\star = 0$.

Adding transport. HCI can reduce copying, but the generated samples may become too concentrated. In the GMM experiment, HCI moves samples away from training points into regions where the learned score pulls them inward much more strongly than the reference score. The constructed HCI field moves samples along reference level sets, so it does not directly remove this inward pull.

To alleviate this problem, we add a transport term guided by the difference between the learned and reference scores. The added term adjusts motion across reference level sets, while HCI continues to move samples along them. We use this transport term as a heuristic correction during sampling, with the network weights fixed. In the GMM experiment, combining the two reduces the concentration seen with HCI alone while keeping copying low. The sampling updates and comparisons are given in SI G.1 and SI G.2.

4.4 HCI SAMPLING

Reverse-process sampling. A score-based diffusion model (Song et al., 2021b) noises data by the VP forward SDE $\mathrm{d}x = f(x,t)\,\mathrm{d}t + g_t\,\mathrm{d}W_t$ with $f = -\frac{1}{2}\beta_t x$ and $g_t = \sqrt{\beta_t}$, whose marginals are $x_t = \sqrt{\bar\alpha_t}\,x_0 + \sigma_t\epsilon$ with $\sigma_t = \sqrt{1-\bar\alpha_t}$, so that $D_t = \frac{1}{2}g_t^2$. It samples by integrating the reverse SDE

$$\mathrm{d}x = \big[f(x,t) - g_t^2\,\hat{\boldsymbol{s}}(x,t)\big]\mathrm{d}t + g_t\,\mathrm{d}\bar W_t \tag{14}$$

backward in time, or the probability-flow ODE with $\frac{1}{2}g_t^2$ in place of g_t^2 and no noise. Practical samplers (ancestral DDPM (Ho et al., 2020), DDIM (Song et al., 2021a), predictor–corrector, Heun) use the noise prediction $\boldsymbol{\varepsilon}(x,t) = -\sigma_t\hat{\boldsymbol{s}}(x,t)$ through the Tweedie estimate $\hat x_0 = (x - \sigma_t\boldsymbol{\varepsilon})/\sqrt{\bar\alpha_t}$ (Efron, 2011), so they depend on the model only through $(x,t)\mapsto\boldsymbol{\varepsilon}(x,t)$, and HCI modifies this map.

Reference. Before sampling, we fit p^{ref} to the training set, or to samples of an early checkpoint of the same run that does not yet copy, from a family whose VP-noised score is closed form. For a Gaussian mixture $\sum_k \pi_k\mathcal{N}(m_k,\Sigma_k)$ fitted by expectation–maximization, or a single Gaussian,

$$s_t^{\mathrm{ref}}(x) = -\sum_k \gamma_{k,t}(x)\,C_{k,t}^{-1}\big(x - \sqrt{\bar\alpha_t}\,m_k\big), \qquad C_{k,t} = \bar\alpha_t\Sigma_k + (1-\bar\alpha_t)I, \tag{15}$$

with $\gamma_{k,t}$ the component responsibilities. Theorem 1 then applies to the fitted reference p^{ref}, not to the unknown data distribution $p^\star$.

Update. At each sampling step, we evaluate the reference score at the current state and construct the HCI field. HCI is applied in a low-noise window while the trained network stays fixed. For the experiments using transport, we also apply the correction described in Section 4.3. The HCI and transport intervals are specified separately for each experiment. The sampling updates are given in SI F.

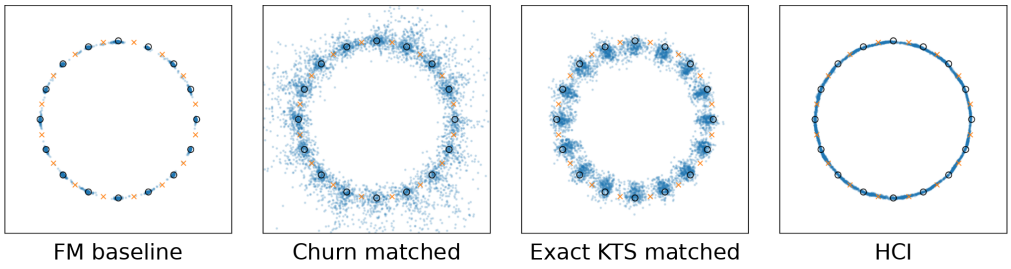


Figure 2: **HCI on the 2D ring.** Endpoints of a frozen, memorized flow-matching model: training points (black circles), held-out midpoints (orange crosses) and samples (blue). At a matched copy fraction, HCI keeps the samples on the circle and moves them into the gaps between training points; churn spreads them off the circle and KTS leaves them near the training points.

5 Experiments

On a 2D ring the field follows from the rotational symmetry of the data and shows visually how a housekeeping current removes memorization. On an analytic GMM and on CelebA in grayscale and color we build the field from a reference fitted to data and compare HCI with the transport correction against churn, KTS and CPSample.

Metrics. For the GMM and CelebA we use the nearest-neighbor gap criterion of Bonnaire et al. (2025): with $d_1(x)$ and $d_2(x)$ the nearest and second-nearest training distances of a sample x, $f_{\mathrm{mem}} = \Pr[d_1(X)/d_2(X) < \gamma_{\mathrm{mem}}]$ with the benchmark threshold $\gamma_{\mathrm{mem}} = 1/3$. For the GMM, whose target is known, we also report $d_1/\ell_{\mathrm{train}}$, with ℓ_{train} the median nearest-neighbor distance within the training set, and the two-sample Kolmogorov–Smirnov (KS) distance between the exact target log-density $\log p^\star$ of the samples and that of fresh target draws, which measures the distribution across level sets of $p^\star$. For CelebA, where the target is unknown, we report FID (Heusel et al., 2017) and copy statistics that a global intensity shift cannot fake (Section 5.3).

5.1 2D ring

The score-based ring in Section 2 examines the learned landscape. For the sampling comparison, we train a separate flow-matching model on the same 16-point ring, so that KTS can be applied in its original setting. We freeze a memorized checkpoint and compare the sampling interventions. The reference is rotationally symmetric, so its level sets are circles and their normal is radial. The field can be written down directly,

$$u_t(x) = w(t)\,\frac{K_2 x}{\|x\|}, \qquad K_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \tag{16}$$

which is tangent to the circles and divergence-free away from the origin, so $\nabla \cdot (p_t^\star u_t) = 0$ for any rotationally symmetric reference and no reference needs to be fitted. We add ηu_t to the frozen flow-matching velocity at low noise (SI F).

The flow-matching baseline is memorized, $f_{\mathrm{mem}} \approx 0.97$ (Figure 2). We tune churn, exact KTS and HCI to the same copy fraction, $f_{\mathrm{mem}} \approx 0.34$–$0.36$, and compare where the remaining samples go using the midpoint occupancy f_{mid}, the fraction of samples near the held-out midpoints (SI E).

- **Churn** spreads the samples off the circle: it reduces memorization by leaving the reference ($f_{\mathrm{mid}} \approx 0.12$).
- **KTS** leaves them clustered around the training points ($f_{\mathrm{mid}} \approx 0.05$).
- **HCI** moves them along the circle into the gaps between training points ($f_{\mathrm{mid}} \approx 0.42$): it removes memorization while preserving the reference.

Reporting both f_{mem} and f_{mid} distinguishes movement along the ring from simply moving samples away from it.

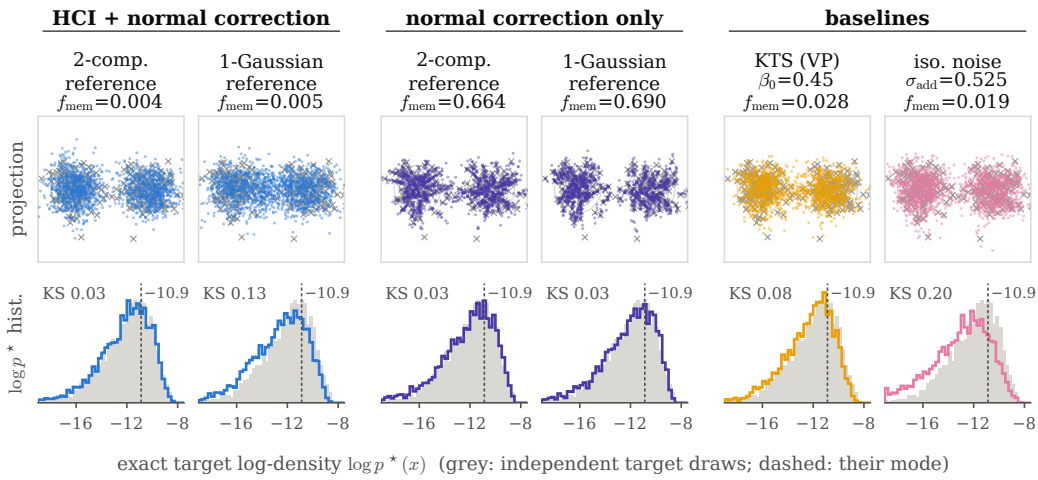


Figure 3: **Analytic GMM.** Top: samples projected onto the mode axis and an orthogonal direction, with training points (×). Bottom: histogram of $\log p^\star$ of the samples against independent target draws (grey), with the KS distance. The dashed line marks the mode of the grey histogram, $\log p^\star \approx -10.9$.

5.2 $\mathbb{R}^8$ GMM

We use the two-component mixture $p^\star = \frac{1}{2}\mathcal{N}(\mu\mathbf{1}, I) + \frac{1}{2}\mathcal{N}(-\mu\mathbf{1}, I)$ in $\mathbb{R}^8$ with $\mu = 1$ and $n = 256$ training points, a memorized score network after 4×10^6 updates ($f_{\text{mem}} = 0.990$), and 4,000 samples per setting with the same initial and sampler noise. HCI uses the constructed field $u_t = -K_{\text{block}} s_t^{\text{ref}}$ of Proposition 3, where $K_{\text{block}} = \text{diag}(K_2, K_2, K_2, K_2)$ rotates coordinate pairs and s_t^{ref} is the closed-form score (15) of a reference fitted to the training points: a two-component Gaussian mixture, or a single Gaussian that sits between the modes. HCI and the normal correction act during the final 30% of the schedule (SI G.1). Our VP adaptation of KTS multiplies the denoising move of each reverse step, from x_t toward the model's posterior mean, by a factor $\zeta \leq 1$ that equals 1 for the first 60% of sampling and then decreases at a rate set by β_0, which damps the last steps that pull samples onto training points. Isotropic noise adds $\sigma_{\text{add}}\,\mathcal{N}(0, I)$ to the final baseline samples.

Figure 3 compares, on the same memorized model, HCI with the normal correction, the correction alone, and two baselines (SI Table 2).

- **HCI + normal correction** removes memorization with either reference ($f_{\text{mem}} = 0.004$ and 0.005). With the two-component reference the log-density histogram matches the target (KS 0.03, at the sampling floor); the single-Gaussian reference leaves a bridge between the modes and a heavier low-density tail (KS 0.13).
- **Normal correction alone** keeps the histogram on the target (KS 0.03) but leaves $f_{\text{mem}} \approx 0.66$–$0.69$, with samples still around training points. **HCI alone** removes copies but contracts the histogram, because between training points the memorized score pulls inward far more strongly than the reference (SI G.2).
- **KTS (VP adaptation)** ($f_{\text{mem}} = 0.028$) stays close to the target (KS 0.08). **Isotropic noise** ($f_{\text{mem}} = 0.019$) shifts the histogram below the target mode (KS 0.20): it mitigates memorization by leaving the target.

5.3 CelebA

We follow the CelebA memorization protocol of Bonnaire et al. (2025): 32×32 grayscale images and a U-Net of base width W trained on n images for τ updates. We select the memorized model $(n, W, \tau) = (512, 32, 2 \times 10^6)$ with $f_{\text{mem}} = 93.4\%$ at FID 18.8 and draw 10,000 DDIM-100 samples per setting from the same initial noise. FID does not detect copies, and pixel f_{mem} can be lowered by changing the global intensity of an image, so we also report the copy fraction after normalizing every image to zero mean and unit variance (std. f_{mem}) and the NN ratio, the median distance to the



Figure 4: **Paired CelebA samples** ($n = 512$, $W = 32$, $\tau = 2 \times 10^6$). For each method, the rows show a sample and its nearest training image, with $r = d_1/d_2$ below (red: copy, $r < 1/3$; green: non-copy). Each column uses the same initial noise for all methods.

nearest training image divided by that to the nearest of 512 held-out images (0.11 for the baseline, 1.00 for held-out faces).

HCI uses a Gaussian reference $p^{\text{ref}} = \mathcal{N}(m, \Sigma)$ whose mean and covariance are estimated from 10,000 samples of an early checkpoint of the same run, before it memorizes (1.05×10^5 updates, $f_{\text{mem}} = 0.14\%$). The field is $u_t = -Ks_t^{\text{ref}}$, where K rotates within the 8 planes spanned by consecutive pairs of the 16 leading eigenvectors of Σ, the main modes of face variation. HCI ($\eta = 0.7$) acts during the final 25% of the schedule, and the normal correction with the same reference at every step (SI H.2). We compare with CPSample at the setting whose FID is closest to that of HCI (confidence threshold $\alpha = 0.49$, scale 100) and with churn ($S_{\text{churn}} = 41$).

Figure 4 follows the same seven initial noises through each method.

- **Baseline:** every sample is indistinguishable from its nearest training image ($r = 0.06$–0.10).
- **HCI + normal correction** turns all seven into non-copies ($r = 0.48$–0.86) and, over the full run, lowers f_{mem} to 4.6% and std. f_{mem} to 3.8% at FID 40.3, without shifting the brightness distribution (SI Table 3). Pose, outline and shading change while the identity is mostly kept in five of the seven columns; the NN ratio of 0.54 lies between a copy (0.11) and an unseen face (1.00).
- **CPSample** at a similar FID (42.5) leaves three of the seven columns as copies and $f_{\text{mem}} = 28.4\%$ (std. 26.9%); in other columns it switches identity or adds grain.
- **Churn** copies a different training image in every column ($r = 0.07$–0.10, $f_{\text{mem}} = 99.7\%$): its added stochasticity redirects copying to other training images.
- **Replication.** With the HCI setting of this run fixed, we reran this configuration with a new seed and two further memorized configurations ($W = 64$; $n = 256$) under a selection rule declared in advance (lowest std. f_{mem} within an FID budget of $+10$ or $+25$) and widened grids for CPSample, KTS and churn. HCI gave the lowest std. f_{mem} in every configuration–budget cell, by a wide margin in five (e.g. 15.3% against 70.0% for CPSample at FID $+5.8$); in the sixth only its weakest setting met the budget and no method removed copies (SI H.3).

At a similar FID, HCI with a fitted reference removes the near-copies that CPSample leaves and churn keeps, by deforming the memorized face rather than generating a new identity.

5.4 CelebA-64 in color

We apply the same approach to a separate experiment on CelebA at 64×64 in color, with three models (M1–M3). We calculate the distance-ratio criterion of Bonnaire et al. (2025), $f_{\text{mem}} = \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}\left[\frac{d_1(x_j)}{d_2(x_j)} < \frac{1}{3}\right]$, where d_1 and d_2 are the nearest and second-nearest Euclidean distances

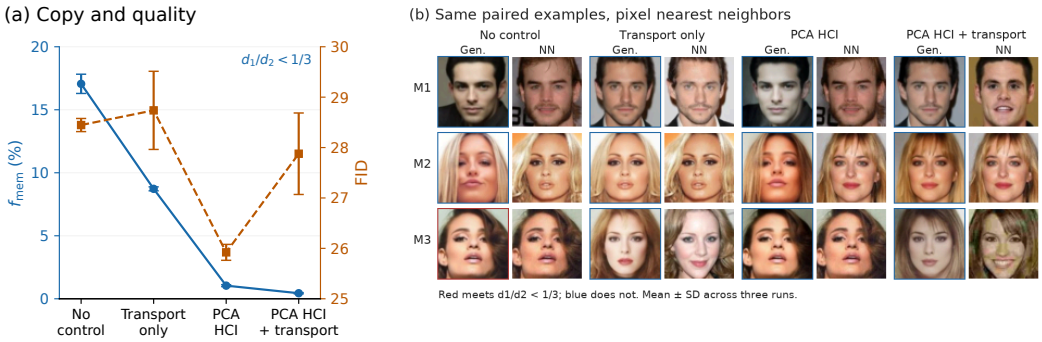


Figure 5: **CelebA-64 sampling controls.** (a) $f_{\mathrm{mem}} = \Pr[d_1/d_2 < 1/3]$ and FID across three models (2,048 samples per model). (b) Paired outputs and nearest training images in pixel space. Red outlines mark samples satisfying the memorization criterion.

in RGB pixel space to the original 128 training images, without horizontal flips. The threshold was fixed before evaluation.

The memorized fraction decreases from $17.06 \pm 0.76\%$ without control to $8.72 \pm 0.16\%$ with transport only, $1.04 \pm 0.07\%$ with PCA HCI, and $0.44 \pm 0.05\%$ with PCA HCI + transport. Thresholds of $1/4$ and $1/2$ give the same ordering.

6 Discussion and Conclusion

We construct HCI from a fitted reference to change sampling paths without retraining the network. When HCI alone causes concentration, we add a heuristic transport correction. On the GMM and grayscale CelebA, HCI with the normal correction removes near-copies almost completely while keeping the structure of the data. On the GMM f_{mem} falls from 0.990 to 0.004 with the exact log-density distribution at the sampling floor, and on CelebA from 93.4% to 4.6% (CPSample at a similar FID: 28.4%), a gain that replicates under a selection rule fixed in advance and is not explained by noise, a brightness shift or copying another image (SI I).

Exact preservation applies to the fitted reference, not to the unknown data distribution. The reference and injection strength must therefore be checked through both copying and sample quality.

AI USE STATEMENT

We used generative AI tools to help implement the methods, including writing and debugging the code and the experiment and evaluation scripts, and to review drafts of the paper. We also used them to search the related literature, to create and modify figures, to draft and edit parts of the text, and to check and format references. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies reduction of near-exact training-example reproduction, which is relevant to privacy and copyright risk. It does not provide a formal privacy guarantee. CelebA is used only as a public low-resolution benchmark; no identity recognition or protected-attribute inference is performed. Qualitative panels compare residual structure and copying behavior, not personal identity.

REPRODUCIBILITY STATEMENT

The SI specifies the fitted references and their closed-form scores, the constructed fields, the transport correction, the integrators and time weights (SI F), the comparison methods and their full grids, the selection rule and success criterion fixed before the CelebA replication runs, all tested settings, the memorization threshold, the metrics and the evaluation sample counts (SI G, H and J). All reported numbers are computed on frozen checkpoints.

REFERENCES

Ricardo Baptista, Agnimitra Dasgupta, Nikola B. Kovachki, Assad Oberai, and Andrew M. Stuart. Memorization and regularization in generative diffusion models. *arXiv preprint arXiv:2501.15785*, 2025.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In *Advances in Neural Information Processing Systems*, 2025.

Andreas Dechant and Shin-ichi Sasa. Entropic bounds on currents in langevin systems. *Physical Review E*, 97(6):062101, 2018.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *arXiv preprint arXiv:2310.02664*, 2023.

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*, volume 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

Kotaro Ikeda, Tomoya Uda, Daisuke Okanohara, and Sosuke Ito. Speed-accuracy relations for diffusion models: Wisdom from nonequilibrium thermodynamics and optimal transport. *Physical Review X*, 15(3):031031, 2025.

Sosuke Ito. Geometric thermodynamics for the Fokker–Planck equation: Stochastic thermodynamic links between information geometry and optimal transport. *Information Geometry*, 7:S441–S483, 2024. doi: 10.1007/s41884-023-00102-3.

Dongjae Jeon, Dueun Kim, and Albert No. Understanding and mitigating memorization in generative models via sharpness of probability landscapes. In *International Conference on Machine Learning*, volume 267, pages 27091–27112, 2025.

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*, volume 35, pages 26565–26577, 2022.

Joshua Kazdan, Hao Sun, Jiaqi Han, Felix Petersen, Frederick Vu, and Stefano Ermon. CPSample: Classifier protected sampling for guarding training data during diffusion. In *International Conference on Learning Representations*, 2025.

Nathan X Kodama and Michael Hinczewski. Thermodynamic performance limits for score-based diffusion models. *arXiv preprint arXiv:2510.06174*, 2025.

Ziyun Li, Huancheng Hu, Soon Hoe Lim, Xuyu Li, Fei Gao, Enmao Diao, Zezhen Ding, Michalis Vazirgiannis, and Henrik Boström. A kinetic-energy perspective of flow matching. In *International Conference on Machine Learning*, 2026. Spotlight.

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

Bao Pham, Gabriel Raya, Matteo Negri, Mohammed J. Zaki, Luca Ambrogioni, and Dmitry Krotov. Memorization to generalization: Emergence of diffusion models from associative memory. *arXiv preprint arXiv:2505.21777*, 2025.

Udo Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics*, 75(12):126001, 2012.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

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*, 2021b.

Benjamin Sterling, Mónica F. Bugallo, and Tom Tirer. Reducing diffusion model memorization with higher order langevin dynamics. *arXiv preprint arXiv:2605.19170*, 2026.

A NOTATION AND CONVENTIONS

This section collects the symbols used throughout the main text and supplementary material. A hat refers to the frozen learned model or a quantity generated by it, and a star refers to the prescribed reference. The score backbone is frozen, so its parameter subscript is suppressed. θ denotes the polar angle.

In Section 3 and SI B, we write $p_t = p_t^\star$, $s_t = s_t^\star$, and $\hat{s}_t = \hat{\boldsymbol{s}}(\cdot, t)$. The scalar coefficient $D(t)$ corresponds to $D_t = D(t)I$. The original EB quantities are $J(v, \Lambda)$, $C[\Lambda]$, $\mathcal{D}_{\mathrm{err}}$, ΔW_{bias}, $\Delta W_{\|}$, and ΔH_{hk}; their definitions are given in Section 3.

Symbol	Meaning
$t \in [0, T]$	Reverse-sampling time. Training progress is reported directly in update counts.
σ_t; σ	Diffusion noise scale along a schedule; in the ring landscape section, σ is one fixed noise level at which the frozen score is evaluated.
$p_t^\star$	Prescribed smooth reference density path to be protected.
$\hat{p}_t$	Uncontrolled marginal generated by the frozen learned sampler.
p_t^η	Density after adding the drift ηu_t to the reference process in Theorem 1.
$b_t^\star$, D_t	Drift and state-independent diffusion tensor of the reference reverse process; its diffusion covariance is $2D_t$.
$v_t^\star$	Reference mean local velocity: $v_t^\star = b_t^\star - D_t \nabla \log p_t^\star$.
$s_t^\star = \nabla \log p_t^\star$	Exact reference score.
$\hat{\boldsymbol{s}}(x, t)$	Frozen vector-valued score-network output. It is not assumed to equal the actual marginal score $\nabla \log \hat{p}_t$.
$\hat{b}_t$, $\hat{v}_t$	Learned reverse drift and learned mean local velocity, respectively; $\hat{v}_t = \hat{b}_t - D_t \nabla \log \hat{p}_t$ and $\partial_t \hat{p}_t + \nabla \cdot (\hat{p}_t \hat{v}_t) = 0$.

Symbol	Ring projection and probability currents
θ, $e_r(\theta)$, $e_\theta(\theta)$	Polar angle, radial unit vector, and counterclockwise tangential unit vector.
$\hat{s}_r$, $\hat{s}_\theta$	Scalar radial and tangential components of the learned vector score: $\hat{s}_r = \hat{\boldsymbol{s}} \cdot e_r$ and $\hat{s}_\theta = \hat{\boldsymbol{s}} \cdot e_\theta$.
$\bar{\hat{s}}_c(r)$	Angular average of score component $c \in \{r, \theta\}$ at radius r.
$\hat{r}$	Learned stable ring radius, defined by $\bar{\hat{s}}_r(\hat{r}) = 0$ and $\partial_r \bar{\hat{s}}_r(\hat{r}) < 0$.
$\hat{E}_t$, $E_t^\star$, ω_t	Learned density energy, reference density energy, and density ratio $\omega_t = \hat{p}_t / p_t^\star$, related exactly by $\hat{E}_t = E_t^\star - \log \omega_t + C_t$.
$E_{\mathrm{rad}}(r)$	One-dimensional radial effective potential obtained from $\bar{\hat{s}}_r$.
$E_{\mathrm{ang}}(\theta)$	Periodic angular effective potential obtained from the zero-mean tangential score on $r = \hat{r}$. It is not a global potential for the full score field.
P_{tr}^p, P_{hk}^p	Weighted Hodge projections onto the gradient component and the component satisfying $\nabla \cdot (pu) = 0$, respectively.
u_t, ηu_t	Drift shape and the actual added drift. The theorem constrains u_t through $\nabla \cdot (p_t^\star u_t) = 0$; η is a scalar strength.
$w(t)$, $\bar{w}(t)$	Time weight of the injection, nonzero only at low noise levels, and its product with the SDE or probability-flow factor.

Symbol	HCI sampler
β_t, $\bar{\alpha}_t$, σ_t	VP noise schedule, signal coefficient and noise scale $\sigma_t = \sqrt{1 - \bar{\alpha}_t}$; $f = -\frac{1}{2}\beta_t x$, $g_t = \sqrt{\beta_t}$, $D_t = \frac{1}{2} g_t^2$.
$\boldsymbol{\varepsilon}$, $\widetilde{\boldsymbol{\varepsilon}}$	Network noise prediction $\boldsymbol{\varepsilon} = -\sigma_t \hat{\boldsymbol{s}}$ and its HCI replacement (97).
p^{ref}, s_t^{ref}, $\varepsilon^{\mathrm{ref}}$	Reference fitted to data, the closed-form score of its VP-noised path, and $\varepsilon^{\mathrm{ref}} = -\sigma_t s_t^{\mathrm{ref}}$.
K, K_2, K_{block}	Constant skew matrix of the constructed field $u_t = -K s_t^{\mathrm{ref}}$; the 2×2 rotation generator; $\mathrm{diag}(K_2, K_2, K_2, K_2)$ on $\mathbb{R}^8$.
n_t, δ_t, δ_t^n, $\delta_t^\perp$	Level-set normal $s_t^\star / \|s_t^\star\|$ (or its fitted counterpart), the score gap $\hat{\boldsymbol{s}} - s_t^\star$, and its normal and tangential parts.
$\hat{\boldsymbol{s}}^{\,c}$	Network score after the transport (normal) correction (96).

Symbol	Evaluation
$d_1(x)$, $d_2(x)$	Nearest and second-nearest distances from generated sample x to the training set.
$r_{\mathrm{gap}}(x)$, γ_{mem}	Gap ratio d_1/d_2 and the benchmark-specific memorization threshold.
f_{mem}	Near-copy fraction: the gap criterion $r_{\mathrm{gap}} < \gamma_{\mathrm{mem}}$ for GMM/CelebA, and the stated endpoint-distance criterion for the ring.
f_{mid}	Fraction of ring endpoints inside neighborhoods of the held-out angular midpoints. It is not the complement of f_{mem} because an endpoint may be near neither a training point nor a midpoint.
ℓ_{train}	Median nearest-neighbor distance between distinct training examples; it is used to normalize d_1.
σ_{add}, σ_{blur}	Standard deviations of isotropic added noise and post-hoc Gaussian blur; neither is the diffusion schedule σ_t.
$\omega_t = \hat{p}_t/p_t^\star$	Learned/reference density ratio.
$\Delta E_t = -\log \omega_t$	Log-density-ratio contribution relative to the reference.
$\mathcal{F}_t[q]$	Relative free energy $D_{\mathrm{KL}}(q\|p_t^\star) = \int q E_t^\star - (-\int q \log q) + \log Z_t$.
Z_t, $\hat{Z}_t$	Normalizers of the reference and learned density energies.
$\mathcal{E}_{\mathrm{tr}}$	Integrated reference-weighted transport error in (93).
τ	Auxiliary relaxation time at a fixed reverse-sampling slice t.
std. f_{mem}	Copy fraction after normalizing every image to zero mean and unit variance.
NN ratio	Median nearest-training distance divided by median nearest-held-out distance (CelebA).

B Entropic Bound and the Two Speed-Limit Branches

We use the notation of Section 3. $p_t = p_t^\star$ is the exact reverse density, v_t its probability velocity, and $D(t) > 0$ the scalar diffusion coefficient. The boundary conditions are stated in SI C. This section gives the EB proof, the endpoint and score-gap bounds, and the transport–housekeeping decomposition.

B.1 Master entropic bound

For a probe field Λ_t, define

$$\Sigma[v] := \int_0^T \mathbb{E}_{p_t}\left[\frac{\|v_t\|^2}{D(t)}\right] \mathrm{d}t, \tag{17}$$

$$J(v,\Lambda) := \int_0^T \mathbb{E}_{p_t}[\Lambda_t \cdot v_t]\,\mathrm{d}t, \tag{18}$$

$$C[\Lambda] := \int_0^T \mathbb{E}_{p_t}[D(t)\|\Lambda_t\|^2]\,\mathrm{d}t. \tag{19}$$

Applying $2|a\cdot b| \le \|a\|^2 + \|b\|^2$ with $a = v_t/\sqrt{D(t)}$ and $b = \sqrt{D(t)}\Lambda_t$, then averaging over p_t and integrating in time, gives

$$|J(v,\Lambda)| \le \frac{\Sigma[v] + C[\Lambda]}{2}. \tag{20}$$

Cauchy–Schwarz over the same space–time measure gives the quadratic EB (Dechant and Sasa, 2018):

$$|J(v,\Lambda)|^2 \le \Sigma[v]C[\Lambda]. \tag{21}$$

For the diffusion's own mean local velocity, $\Sigma[v]$ is its entropy-production functional in the convention used here (Ito, 2024). The quadratic inequality also holds for an arbitrary velocity field, but its squared norm should not then be identified with the total entropy production of a different process.

B.2 Endpoint error: Wasserstein distance

To isolate the effect of an initial density mismatch, assume that the exact and estimated densities evolve with the same probability velocity:

$$\partial_t p_t + \nabla\cdot(p_t v_t) = 0, \qquad \partial_t \hat{p}_t + \nabla\cdot(\hat{p}_t v_t) = 0. \tag{22}$$

Define

$$\delta p_t := p_t - \hat{p}_t, \qquad r_t := \frac{\delta p_t}{p_t}. \tag{23}$$

Then $\partial_t \delta p_t = -\nabla \cdot (\delta p_t v_t)$ and $\partial_t r_t + v_t \cdot \nabla r_t = 0$. Integrating by parts shows that the Pearson divergence is constant:

$$R_0 := \int \frac{(p_0 - \hat{p}_0)^2}{p_0} \, \mathrm{d}x = \mathbb{E}_{p_0}[r_0^2] = \mathbb{E}_{p_t}[r_t^2]. \tag{24}$$

Let ψ_t be a Kantorovich witness for $W_1(p_t, \hat{p}_t)$, with $\|\nabla \psi_t\| \le 1$, and choose

$$\Lambda_t(x) := r_t(x) \nabla \psi_t(x). \tag{25}$$

For a fixed test function ψ, the continuity equations give

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\mathbb{E}_{p_t}[\psi] - \mathbb{E}_{\hat{p}_t}[\psi] \right) = \int \delta p_t \nabla \psi \cdot v_t \, \mathrm{d}x. \tag{26}$$

Evaluating at an optimal witness gives, at differentiability times,

$$\partial_t W_1(p_t, \hat{p}_t) = \mathbb{E}_{p_t}[r_t \nabla \psi_t \cdot v_t]. \tag{27}$$

The corresponding one-sided statement is used at exceptional times. With

$$j_t(v, \Lambda) := \mathbb{E}_{p_t}[\Lambda_t \cdot v_t], \quad \dot{\Sigma}(t) := \mathbb{E}_{p_t}\left[\frac{\|v_t\|^2}{D(t)} \right], \quad c_t[\Lambda] := D(t) \mathbb{E}_{p_t} \|\Lambda_t\|^2, \tag{28}$$

the instantaneous EB reads $|j_t(v, \Lambda)|^2 \le \dot{\Sigma}(t) c_t[\Lambda]$. Since

$$c_t[\Lambda] = D(t) \mathbb{E}_{p_t}[r_t^2 \|\nabla \psi_t\|^2] \le D(t) R_0, \tag{29}$$

we obtain

$$\frac{|\partial_t W_1(p_t, \hat{p}_t)|^2}{R_0} \le D(t) \dot{\Sigma}(t). \tag{30}$$

Integrating over time and applying Cauchy–Schwarz yields

$$\frac{(\Delta W_1)^2}{T R_0} \le \int_0^T D(t) \dot{\Sigma}(t) \, \mathrm{d}t, \qquad \Delta W_1 := W_1(p_T, \hat{p}_T) - W_1(p_0, \hat{p}_0). \tag{31}$$

This is the endpoint speed–accuracy relation (Ikeda et al., 2025). It bounds the change in distributional error caused by an initial mismatch under the shared velocity. If $R_0 = 0$, both densities remain equal under this assumption, so the quotient form is unnecessary.

Different learned and reference velocities. If $\hat{p}_t$ instead evolves with its own velocity $\hat{v}_t$, the witness derivative has an additional term:

$$\partial_t W_1(p_t, \hat{p}_t) = \mathbb{E}_{p_t}[r_t \nabla \psi_t \cdot v_t] + \int \hat{p}_t \nabla \psi_t \cdot (v_t - \hat{v}_t) \, \mathrm{d}x. \tag{32}$$

Its absolute value is bounded by

$$\left| \int \hat{p}_t \nabla \psi_t \cdot (v_t - \hat{v}_t) \, \mathrm{d}x \right| \le \left(D(t) \mathbb{E}_{\hat{p}_t}\left[\frac{\|v_t - \hat{v}_t\|^2}{D(t)} \right] \right)^{1/2}. \tag{33}$$

The Pearson divergence need not remain constant in this case. Thus (31) is not applied unchanged to unequal velocities. The assumption on the learned transport in (93) remains a separate assumption.

B.3 Score-gap branch: cumulative transport bias

Let $s_t = \nabla \log p_t = s_t^\star$ and $\hat{s}_t = \hat{\boldsymbol{s}}(\cdot, t)$ be the reference and learned scores. Choose $\Lambda_t = \hat{s}_t - s_t$ in EB. Then

$$\left\{ \int_0^T \mathbb{E}_{p_t}[(\hat{s}_t - s_t) \cdot v_t] \, \mathrm{d}t \right\}^2 \le \Sigma[v] \int_0^T D(t) \mathbb{E}_{p_t} \|\hat{s}_t - s_t\|^2 \, \mathrm{d}t. \tag{34}$$

Define the accumulated bias and time-averaged error diffusivity by

$$\Delta W_{\mathrm{bias}} := \int_0^T \mathbb{E}_{p_t}[(\hat{s}_t - s_t) \cdot v_t]\,\mathrm{d}t, \tag{35}$$

$$\mathcal{D}_{\mathrm{err}} := \frac{1}{T}\int_0^T D(t)\mathbb{E}_{p_t}\|\hat{s}_t - s_t\|^2\,\mathrm{d}t. \tag{36}$$

The speed–accuracy form is

$$\frac{(\Delta W_{\mathrm{bias}})^2}{T\mathcal{D}_{\mathrm{err}}} \le \Sigma[v]. \tag{37}$$

In the convention used for the likelihood score-mismatch term (Kodama and Hinczewski, 2025),

$$\Delta\mathrm{NLL}_{\mathrm{gap}} = \frac{1}{2}\int_0^T D(t)\mathbb{E}_{p_t}\|\hat{s}_t - s_t\|^2\,\mathrm{d}t = \frac{T\mathcal{D}_{\mathrm{err}}}{2}. \tag{38}$$

Thus the score-gap EB retains the component of the error along the reverse current, in addition to the total squared score mismatch.

Meaning of the velocity error. We retain the score-based definition

$$\delta v_t := D(t)(\hat{s}_t - s_t), \qquad T\mathcal{D}_{\mathrm{err}} = \int_0^T \mathbb{E}_{p_t}\left[\frac{\|\delta v_t\|^2}{D(t)}\right]\mathrm{d}t. \tag{39}$$

Under the shared-drift comparison, this is the difference between the two score-based velocity fields:

$$(b_t - D(t)s_t) - (b_t - D(t)\hat{s}_t) = \delta v_t. \tag{40}$$

The shared-drift condition is needed for this interpretation. It is not automatically the difference between the actual reference and learned mean local velocities, because the learned reverse drift can differ and the network output need not equal $\nabla \log \hat{p}_t$. The bounds above require only the definition of δv_t from the score error.

If $\hat{s}_t = \nabla \log \hat{p}_t$, then

$$\hat{s}_t - s_t = \nabla \log \frac{\hat{p}_t}{p_t^\star} = -\nabla \Delta E_t. \tag{41}$$

This connects the score-gap observable to the energy difference without assuming that a general network output is the gradient of a global potential.

B.4 Weighted Hodge decomposition and the transport minimizer

For a density p and a state-independent positive-definite tensor D, the spatial inner product is $\langle a, b\rangle_{p,D} = \int p\, a^\top D^{-1} b\,\mathrm{d}x$. A state observable f tests density change through its gradient: at a fixed time,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_{p_t}[f]\Big|_{\mathrm{current}} = \int p_t \nabla f \cdot v_t\,\mathrm{d}x. \tag{42}$$

Any field u orthogonal to all such gradient probes satisfies $\nabla\cdot(pu) = 0$ in the weak sense. The following minimization makes the complementary transport field explicit; the constraint comes from the continuity equation. Fix a positive density p, a positive-definite tensor D, and a raw velocity field a. Among all velocities producing the same instantaneous density source as a, consider

$$\min_v \frac{1}{2}\int p\, v^\top D^{-1} v\,\mathrm{d}x \quad \text{s.t.} \quad \nabla\cdot(pv) = \nabla\cdot(pa). \tag{43}$$

The Lagrangian can be written

$$\mathcal{L}(v,\chi) = \frac{1}{2}\int p\, v^\top D^{-1} v\,\mathrm{d}x + \int \chi\,[\nabla\cdot(pv) - \nabla\cdot(pa)]\,\mathrm{d}x. \tag{44}$$

After integration by parts, its variation is $\delta_v\mathcal{L} = \int p(D^{-1}v - \nabla\chi)\cdot\delta v\,\mathrm{d}x$. At the minimum, $D^{-1}v - \nabla\chi = 0$, so $v = D\nabla\chi$, with

$$\nabla\cdot(pD\nabla\chi) = \nabla\cdot(pa). \tag{45}$$

Under the usual coercivity or weighted-Poincaré assumptions, χ is unique up to an additive constant. Define

$$P^p_{\mathrm{tr}} a = D\nabla\chi, \qquad P^p_{\mathrm{hk}} a = a - D\nabla\chi. \tag{46}$$

Then $\nabla\cdot(pP^p_{\mathrm{hk}}a) = 0$ and

$$\left\langle P^p_{\mathrm{tr}} a, P^p_{\mathrm{hk}} a \right\rangle_{p,D} = \int p\,(D\nabla\chi)^\top D^{-1} P^p_{\mathrm{hk}} a \,\mathrm{d}x \tag{47}$$

$$= \int p\, P^p_{\mathrm{hk}} a \cdot \nabla\chi \,\mathrm{d}x = 0. \tag{48}$$

Therefore

$$\|a\|^2_{p,D} = \|P^p_{\mathrm{tr}} a\|^2_{p,D} + \|P^p_{\mathrm{hk}} a\|^2_{p,D}. \tag{49}$$

This is a density-relative decomposition: $P^p_{\mathrm{hk}}a$ and $P^q_{\mathrm{hk}}a$ generally differ. The term housekeeping refers to the divergence-free probability current $pP^p_{\mathrm{hk}}a$, not to a velocity field independently of its density weight.

For any other admissible field $v = D\nabla\chi + u$ with $\nabla\cdot(pu) = 0$,

$$\|v\|^2_{p,D} = \|D\nabla\chi\|^2_{p,D} + \|u\|^2_{p,D}. \tag{50}$$

This proves minimum norm and uniqueness of the transport component (in the weighted function space), while leaving an entire class of possible housekeeping fields. Integrating this orthogonality in time gives the entropy-production split in (6).

For the scalar coefficient used in Section 3, set $D = D(t)I$. The squared spatial norm is $\|a\|^2_{p,D} = \langle a, a\rangle_{p,D}$; integrating it in time gives the inner product $\langle\cdot,\cdot\rangle$ defined after (6).

B.5 Longitudinal bias and housekeeping residual

The decomposition is

$$v_t = v_t^{\mathrm{tr}} + v_t^{\mathrm{hk}}, \qquad \nabla\cdot(p_t v_t^{\mathrm{hk}}) = 0. \tag{51}$$

Using the time-integrated inner product $\langle\cdot,\cdot\rangle$ defined after (6), define

$$\Delta W_{\|} := \langle \delta v, v^{\mathrm{tr}}\rangle = \int_0^T \mathbb{E}_{p_t}\left[\frac{\delta v_t \cdot v_t^{\mathrm{tr}}}{D(t)}\right] \mathrm{d}t, \tag{52}$$

$$\Delta H_{\mathrm{hk}} := \langle \delta v, v^{\mathrm{hk}}\rangle = \int_0^T \mathbb{E}_{p_t}\left[\frac{\delta v_t \cdot v_t^{\mathrm{hk}}}{D(t)}\right] \mathrm{d}t, \tag{53}$$

$$\Sigma_{\mathrm{tr}} := \int_0^T \mathbb{E}_{p_t}\left[\frac{\|v_t^{\mathrm{tr}}\|^2}{D(t)}\right] \mathrm{d}t, \qquad \Sigma_{\mathrm{hk}} := \int_0^T \mathbb{E}_{p_t}\left[\frac{\|v_t^{\mathrm{hk}}\|^2}{D(t)}\right] \mathrm{d}t. \tag{54}$$

Linearity gives $\Delta W_{\mathrm{bias}} = \Delta W_{\|} + \Delta H_{\mathrm{hk}}$. Applying EB to the same score probe and each current component gives

$$(\Delta W_{\|})^2 \le \Sigma_{\mathrm{tr}} T\mathcal{D}_{\mathrm{err}}, \qquad (\Delta H_{\mathrm{hk}})^2 \le \Sigma_{\mathrm{hk}} T\mathcal{D}_{\mathrm{err}}. \tag{55}$$

For $\mathcal{D}_{\mathrm{err}} > 0$, these are equivalent to

$$\frac{(\Delta W_{\|})^2}{T\mathcal{D}_{\mathrm{err}}} \le \Sigma_{\mathrm{tr}}, \qquad \frac{(\Delta H_{\mathrm{hk}})^2}{T\mathcal{D}_{\mathrm{err}}} \le \Sigma_{\mathrm{hk}}. \tag{56}$$

If $\mathcal{D}_{\mathrm{err}} = 0$, both biases vanish and the product form remains valid.

The two components are orthogonal. When both have positive norm, Bessel's inequality applied to $v^{\mathrm{tr}}/\sqrt{\Sigma_{\mathrm{tr}}}$ and $v^{\mathrm{hk}}/\sqrt{\Sigma_{\mathrm{hk}}}$ yields

$$T\mathcal{D}_{\mathrm{err}} \ge \frac{(\Delta W_{\|})^2}{\Sigma_{\mathrm{tr}}} + \frac{(\Delta H_{\mathrm{hk}})^2}{\Sigma_{\mathrm{hk}}}. \tag{57}$$

A term with zero component norm is omitted, since its numerator is also zero. The longitudinal bias is a path-integrated current observable, not a Wasserstein distance between endpoints. The housekeeping residual measures the corresponding projection onto the divergence-free component.

For an exact gradient score gap $\hat{s}_t - s_t = \nabla f_t$, integration by parts gives

$$\Delta H_{\mathrm{hk}} = \int_0^T \int p_t v_t^{\mathrm{hk}} \cdot \nabla f_t \,\mathrm{d}x\,\mathrm{d}t = -\int_0^T \int f_t \nabla\cdot(p_t v_t^{\mathrm{hk}})\,\mathrm{d}x\,\mathrm{d}t = 0. \tag{58}$$

A general vector-valued network score error need not be a gradient, so this projection can be nonzero. It is not itself a measure of memorization.

B.6 HOUSEKEEPING-CURRENT INJECTION

The current decomposition is relative to a specified density. An added drift ηu_t with

$$\nabla\cdot(p_t^\star u_t) = 0 \tag{59}$$

leaves the reference path unchanged under the assumptions of Theorem 1. For the learned density $\hat{p}_t = p_t^\star \omega_t$, however,

$$-\nabla\cdot(\hat{p}_t u_t) = -p_t^\star u_t \cdot \nabla\omega_t. \tag{60}$$

The same field can therefore change the learned sampling paths and endpoint probabilities. HCI adds such a field; it does not estimate and cancel the baseline v_t^{hk}. The bounds separate transport and housekeeping contributions but do not determine the sign of the change in copying. That effect is tested in the experiments.

C WEAK FORMULATION AND PROOFS OF THE EXACT HCI PROPERTIES

This section records the assumptions and proof of Theorem 1, the density-source proof of Proposition 2, and its implication for copying. The proof of Proposition 1 is in SI D.

C.1 ADMISSIBLE DENSITIES, CURRENTS, AND BOUNDARIES

For each sampling time $t \in [0, T]$, let $p_t^\star$ be a strictly positive probability density on either a periodic domain, $\mathbb{R}^d$ with sufficient decay, or a bounded domain with no-flux boundary conditions. We take the diffusion tensor D_t to be state-independent and uniformly positive definite. A velocity field u_t is square-integrable when

$$\int_0^T \int p_t^\star(x) u_t(x)^\top D_t^{-1} u_t(x)\,\mathrm{d}x\,\mathrm{d}t < \infty. \tag{61}$$

The weighted divergence-free condition is understood distributionally: for every smooth compactly supported or boundary-compatible test function ψ,

$$\nabla\cdot(p_t^\star u_t) = 0 \quad\Longleftrightarrow\quad \int p_t^\star u_t \cdot \nabla\psi\,\mathrm{d}x = 0. \tag{62}$$

On a bounded domain we additionally require $(p_t^\star u_t)\cdot n = 0$. The reference and controlled Fokker–Planck equations are assumed to be well posed in the chosen weak solution class. Uniformly elliptic diffusions with standard drift regularity satisfy this assumption; deterministic probability-flow equations require the corresponding uniqueness condition for the continuity equation.

C.2 WEIGHTED HODGE PROJECTION USED IN THE THEOREM

The construction and variational proof are in SI B.4. In particular, $P_{\mathrm{tr}}^p u = D\nabla\chi$, where $\nabla\cdot(pD\nabla\chi) = \nabla\cdot(pu)$, and $P_{\mathrm{hk}}^p u = u - P_{\mathrm{tr}}^p u$. The boundary and coercivity assumptions are those stated above and in SI B.4.

C.3 PROOF THAT THE CONDITION HOLDS FOR ANY CONTROL STRENGTH

Following Ito (2024), the reference mean local velocity is $v_t^\star = b_t^\star - D_t\nabla\log p_t^\star$. Its probability current is $p_t^\star v_t^\star = b_t^\star p_t^\star - D_t\nabla p_t^\star$, so

$$\partial_t p_t^\star = -\nabla\cdot(b_t^\star p_t^\star - D_t\nabla p_t^\star) = -\nabla\cdot(p_t^\star v_t^\star)\,. \tag{63}$$

Adding the prescribed drift ηu_t while retaining the same diffusion tensor gives

$$\partial_t p_t^\eta = -\nabla\cdot[(b_t^\star + \eta u_t)p_t^\eta - D_t\nabla p_t^\eta]\,, \qquad p_0^\eta = p_0^\star. \tag{64}$$

Necessity. Assume $p_t^\eta = p_t^\star$ for every t. Substituting the same density into (63) and (64) and subtracting yields

$$\eta\,\nabla\cdot(p_t^\star u_t) = 0. \tag{65}$$

Since $\eta \neq 0$, $\nabla\cdot(p_t^\star u_t) = 0$ for almost every time. This is the weak-solution meaning of "at each sampling time" in the main text: exceptional time points of measure zero do not affect the time-integrated density evolution.

Sufficiency. Conversely, suppose $\nabla\cdot(p_t^\star u_t)=0$. Inserting $p_t^\star$ into the controlled equation makes the added source vanish, so $p_t^\star$ is a weak solution of the controlled equation with the correct initial density. Uniqueness gives $p_t^\eta = p_t^\star$.

Equivalence with zero transport projection. If $P_{\mathrm{tr}}^{p_t^\star} u_t = 0$, the weighted divergence is zero by construction. Conversely, if $\nabla\cdot(p_t^\star u_t)=0$, the right-hand side of the weighted Poisson equation is zero. Under the stated boundary and uniqueness conditions, χ_t is constant, so $P_{\mathrm{tr}}^{p_t^\star} u_t = D_t\nabla\chi_t = 0$.

For $\eta = 0$, the controlled and reference processes agree for every u_t, so the converse must be stated for a nonzero strength. The condition does not depend on η. If a fixed field preserves the reference at one nonzero value of η, necessity gives $\nabla\cdot(p_t^\star u_t)=0$. The same field therefore preserves the path for every value of η for which the controlled equation remains well posed.

C.4 Orthogonality to the score is not sufficient

For the reference path and candidate intervention,

$$\nabla\cdot(p_t^\star u_t) = p_t^\star\left(\nabla\cdot u_t + u_t\cdot s_t^\star\right). \tag{66}$$

Thus $u_t\cdot s_t^\star = 0$ would only prevent motion across reference-score contours; it would not prevent compression or expansion along them. Reference preservation also requires $\nabla\cdot u_t = 0$ in this score-orthogonal case. A projection against the network output $\hat{\boldsymbol{s}}$, which need not equal either $s_t^\star$ or the actual marginal score $\nabla\log\hat{p}_t$, controls neither term. The constructed fields $u_t = -Ks_t^{\mathrm{ref}}$ satisfy both conditions for the fitted reference (Proposition 3). The ring uses geometric tangency to a rotationally symmetric reference; it need not be orthogonal to the memorized network score, whose angular component is generally nonzero.

C.5 Learned-density response and probability flux near training samples

For any admissible q, expand the added current as

$$-\nabla\cdot(qu_t) = -\nabla\cdot\left[p_t^\star u_t\frac{q}{p_t^\star}\right] \tag{67}$$

$$= -\frac{q}{p_t^\star}\nabla\cdot(p_t^\star u_t) - p_t^\star u_t\cdot\nabla\frac{q}{p_t^\star} \tag{68}$$

$$= -p_t^\star u_t\cdot\nabla\frac{q}{p_t^\star}. \tag{69}$$

Taking $q=\hat{p}_t$ proves Proposition 2. The expression is the instantaneous source contributed by the added drift at the indicated density, not a finite-time formula for the controlled marginal. Multiplication by η gives the actual source.

For a fixed training neighborhood $\mathcal{B}_k$, its added mass rate is

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{B}_k} q\,\mathrm{d}x\bigg|_{u_t} = -\eta\int_{\partial\mathcal{B}_k} qu_t\cdot n\,\mathrm{d}S. \tag{70}$$

Reference preservation does not determine the sign of this boundary integral when $q\neq p_t^\star$.

At each time, decompose the uncontrolled learned mean local velocity with respect to the learned marginal itself,

$$\hat{v}_t = P_{\mathrm{tr}}^{\hat{p}_t}\hat{v}_t + P_{\mathrm{hk}}^{\hat{p}_t}\hat{v}_t, \qquad \nabla\cdot\left(\hat{p}_t P_{\mathrm{hk}}^{\hat{p}_t}\hat{v}_t\right) = 0. \tag{71}$$

The continuity equation then reduces to

$$\partial_t\hat{p}_t = -\nabla\cdot\left(\hat{p}_t P_{\mathrm{tr}}^{\hat{p}_t}\hat{v}_t\right). \tag{72}$$

Let $\mathcal{B}_k$ be a fixed neighborhood of training sample x_k during the low-noise part of reverse sampling. The divergence theorem gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{B}_k}\hat{p}_t\,\mathrm{d}x = -\int_{\partial\mathcal{B}_k}\hat{p}_t P_{\mathrm{tr}}^{\hat{p}_t}\hat{v}_t\cdot n\,\mathrm{d}S. \tag{73}$$

Thus inward probability flux across $\partial\mathcal{B}_k$ increases the probability mass near a training sample. The component satisfying $\nabla\cdot(\hat{p}_t u_t) = 0$ contributes zero to this derivative. This does not imply that every gradient field increases copying: the sign is set by the boundary flux. In the ring example, the angular score near a training point directs trajectories toward that point.

Now suppose that $\nabla\cdot(p_t^\star u_t) = 0$ along the reference path. Then

$$P_{\mathrm{tr}}^{p_t^\star} u_t = 0, \qquad P_{\mathrm{tr}}^{\hat{p}_t} u_t \text{ is generally nonzero,} \tag{74}$$

and

$$-\nabla\cdot(\hat{p}_t u_t) = -\nabla\cdot\left(\hat{p}_t P_{\mathrm{tr}}^{\hat{p}_t} u_t\right). \tag{75}$$

The word "housekeeping" is therefore relative to a specified density. A field satisfying $\nabla\cdot(p_t^\star u_t) = 0$ can still change $\hat{p}_t$. With a suitable sign, HCI can reduce inward probability flux near training samples while leaving the reference density unchanged. The constructed fields of the GMM and CelebA experiments satisfy the condition exactly for the fitted reference p^{ref}, not for the unknown $p^\star$; relative to $p^\star$ they may contain a transport component.

C.6 Exact preservation of the learned path cannot reduce its copy fraction

If $\nabla\cdot(\hat{p}_t u_t) = 0$ along the uncontrolled learned path and the learned and controlled equations share their initial density and diffusion, apply Theorem 1 with $\hat{p}_t$ in place of $p_t^\star$. The endpoint law is unchanged. For the fixed copy set $\mathcal{M}$ determined by the training set and its evaluation threshold,

$$\Pr_{\text{controlled}}[X_T \in \mathcal{M}] = \int_{\mathcal{M}} \hat{p}_T(x)\,\mathrm{d}x = \Pr_{\text{baseline}}[X_T \in \mathcal{M}]. \tag{76}$$

Sample paths and finite-sample estimates may differ, but the population copy probability does not. The relevant freedom is therefore preservation of a different, prescribed reference, not preservation of the memorized marginal itself.

D Free energy, KL relaxation, and the learned-transport assumption

The algebra below expands Sections 2 and 4.2. The energy/free-energy identities and Proposition 1 do not require a small learned-transport error. That assumption is introduced only when interpreting the frozen memorized sampler.

D.1 Energy, entropy, and relative free energy

For fixed t, let

$$p_t^\star = Z_t^{-1} e^{-E_t^\star}, \qquad \hat{p}_t = \hat{Z}_t^{-1} e^{-\hat{E}_t}, \qquad \omega_t = \frac{\hat{p}_t}{p_t^\star}. \tag{77}$$

Then $\log\omega_t = E_t^\star - \hat{E}_t + \log Z_t - \log\hat{Z}_t$, hence

$$\hat{E}_t = E_t^\star - \log\omega_t + C_t, \qquad C_t = \log Z_t - \log\hat{Z}_t. \tag{78}$$

Additive energy constants change the normalizers but not either density. A local excess of learned probability relative to the reference produces a negative $\Delta E_t = -\log\omega_t$. Calling it a narrow training-point well additionally requires the localized geometry observed in the trained-ring experiment; that geometry does not follow merely from the density-ratio identity.

For a normalized density q with finite terms,

$$D_{\mathrm{KL}}(q\|p_t^\star) = \int q\log q\,\mathrm{d}x - \int q\log p_t^\star\,\mathrm{d}x \tag{79}$$

$$= \int qE_t^\star\,\mathrm{d}x + \int q\log q\,\mathrm{d}x + \log Z_t. \tag{80}$$

This is the dimensionless relative free energy: mean reference energy minus entropy, measured relative to the reference equilibrium free energy $-\log Z_t$. It is nonnegative and minimized at $q = p_t^\star$.

Applying the same calculation to the learned landscape gives $D_{\mathrm{KL}}(q\|\hat{p}_t)$, whose minimizer is instead the memorized density. This explains why the reference must remain explicit.

With the usual unconstrained representative of the variational derivative,

$$\frac{\delta \mathcal{F}_t}{\delta q} = \log q + 1 + E_t^\star + \log Z_t = \log \frac{q}{p_t^\star} + 1. \tag{81}$$

The additive constant has no effect on its spatial gradient. These identities do not assume that the network output is an exact density score. Only when $\hat{\boldsymbol{s}} = \nabla \log \hat{p}_t$ may one identify $\hat{\boldsymbol{s}} - s_t^\star = -\nabla \Delta E_t$ with the density-energy gradient.

D.2 KL RELAXATION AT A FIXED SAMPLING TIME

At the reverse-time slice where HCI is applied, hold the reference density fixed. Let $\tau \geq 0$ denote the auxiliary relaxation time and let q_τ start from the learned marginal, $q_0 = \hat{p}_t$:

$$p_t^\star(x) = Z_t^{-1} e^{-E_t^\star(x)}. \tag{82}$$

Along this relaxation, denote the KL divergence from the reference by

$$\mathcal{F}_t[q_\tau] := D_{\mathrm{KL}}(q_\tau \| p_t^\star) = \int q_\tau \log \frac{q_\tau}{p_t^\star} \, \mathrm{d}x. \tag{83}$$

Its variational derivative is $\delta \mathcal{F}_t / \delta q_\tau = \log(q_\tau / p_t^\star) + 1$. The velocity that decreases this KL divergence is

$$v_{t,\mathrm{eq}}^{\mathrm{tr}}[q_\tau] = -D_t \nabla \frac{\delta \mathcal{F}_t}{\delta q_\tau} = -D_t \nabla \log \frac{q_\tau}{p_t^\star}. \tag{84}$$

The relaxation equation $\partial_\tau q_\tau = -\nabla \cdot (q_\tau v_{t,\mathrm{eq}}^{\mathrm{tr}})$ conserves mass. Since $p_t^\star$ is fixed, differentiating the KL divergence and integrating by parts gives

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\tau} \mathcal{F}_t[q_\tau] &= \int \left(\log \frac{q_\tau}{p_t^\star} + 1 \right) \partial_\tau q_\tau \, \mathrm{d}x \\ &= \int q_\tau v_{t,\mathrm{eq}}^{\mathrm{tr}}[q_\tau] \cdot \nabla \log \frac{q_\tau}{p_t^\star} \, \mathrm{d}x, \end{aligned}$$

where the $+1$ term vanishes because $\int \partial_\tau q_\tau \, \mathrm{d}x = 0$. Substituting Equation (84) yields

$$\frac{\mathrm{d}}{\mathrm{d}\tau} \mathcal{F}_t[q_\tau] = -\int q_\tau \left(\nabla \log \frac{q_\tau}{p_t^\star} \right)^\top D_t \nabla \log \frac{q_\tau}{p_t^\star} \, \mathrm{d}x = -\| v_{t,\mathrm{eq}}^{\mathrm{tr}}[q_\tau] \|_{q_\tau, D_t}^2 \leq 0. \tag{85}$$

Thus this velocity decreases the KL divergence and is zero when $q_\tau = p_t^\star$.

D.3 KL NEUTRALITY AND THE FULL RELAXATION BALANCE

Let u_t satisfy $\nabla \cdot (p_t^\star u_t) = 0$ and consider

$$\partial_\tau q_\tau = \nabla \cdot \left[q_\tau D_t \nabla \log \frac{q_\tau}{p_t^\star} \right] - \eta \nabla \cdot (q_\tau u_t). \tag{86}$$

At $q_\tau = p_t^\star$ the first term in Equation (86) vanishes and the added probability current has zero divergence, so $p_t^\star$ is invariant for every admissible η. Moreover,

$$\left. \frac{\mathrm{d}}{\mathrm{d}\tau} \mathcal{F}_t[q_\tau] \right|_{u_t} = \eta \int q_\tau u_t \cdot \nabla \log \frac{q_\tau}{p_t^\star} \, \mathrm{d}x \tag{87}$$

$$= \eta \int p_t^\star u_t \cdot \nabla \left(\frac{q_\tau}{p_t^\star} \right) \mathrm{d}x \tag{88}$$

$$= -\eta \int \frac{q_\tau}{p_t^\star} \nabla \cdot (p_t^\star u_t) \, \mathrm{d}x = 0. \tag{89}$$

Hence the full KL-divergence balance remains

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mathcal{F}_t[q_\tau] = -\int q_\tau \left(\nabla \log \frac{q_\tau}{p_t^\star}\right)^\top D_t \nabla \log \frac{q_\tau}{p_t^\star}\,\mathrm{d}x. \tag{90}$$

The cancellation above holds for every admissible q, so it proves Proposition 1, not just a linearization about $p_t^\star$. Under pure advection $\partial_\tau q = -\eta\nabla\cdot(qu_t)$, the KL divergence is conserved. In the combined relaxation, the current makes no direct contribution to dissipation but changes q_τ; it can therefore affect later relaxation and transition times. No monotone improvement with η is implied. The auxiliary time τ must not be identified with the time-dependent learned reverse process.

The same statement has an operator form. Define $A_{u_t} f = u_t \cdot \nabla f$. For compatible boundaries,

$$\langle f, A_{u_t} g\rangle_{p_t^\star} = \int p_t^\star f u_t \cdot \nabla g\,\mathrm{d}x \tag{91}$$

$$= -\int p_t^\star g u_t \cdot \nabla f\,\mathrm{d}x = -\langle A_{u_t} f, g\rangle_{p_t^\star}, \tag{92}$$

so A_{u_t} is formally skew-adjoint on compatible test functions in $L^2(p_t^\star)$. An operator-level skew-adjoint statement additionally requires the appropriate domain and closed realization. This is the standard operator description of nonreversible sampling: one part decreases the KL divergence, while the added part changes trajectories without changing the invariant density.

D.4 Assumption on the density-changing error in the learned velocity

Comparing the learned mean local velocity $\hat{v}_t$ with the reference velocity $v_t^\star = b_t^\star - D_t\nabla \log p_t^\star$, we measure the error in the part that can change $p_t^\star$ by

$$\mathcal{E}_{\mathrm{tr}} := \int_0^T \left\| P_{\mathrm{tr}}^{p_t^\star}(\hat{v}_t - v_t^\star)\right\|_{p_t^\star, D_t}^2 \mathrm{d}t. \tag{93}$$

We assume this quantity is small on the scales relevant to the experiment. This is not a consequence of HCI or of any temporal ordering during training. The retained annular geometry and acceptable sample-quality metrics motivate asking whether the assumption is useful; neither observation measures or bounds $\mathcal{E}_{\mathrm{tr}}$. In particular, a low FID or a stable ring radius does not establish closeness of the corresponding transport fields.

Under this assumption, HCI is not used to correct the density-changing part of the learned drift. It adds a probability current that preserves $p_t^\star$. For the learned density,

$$-\nabla\cdot(\hat{p}_t u_t) = -p_t^\star u_t \cdot \nabla\left(\frac{\hat{p}_t}{p_t^\star}\right), \tag{94}$$

so the same current can change endpoint probabilities near training samples even though it leaves the reference density invariant.

An exact HCI field satisfies $P_{\mathrm{tr}}^{p_t^\star} u_t = 0$ and thus supplies no direct addition to the reference-weighted transport component. If the transport error is large, reference preservation alone offers no guarantee of correcting the learned distribution. A field constructed from a fitted reference is exact for p^{ref} but may contain transport relative to $p^\star$. Section 4.3 therefore corrects the normal component of the score gap directly; on the GMM this gap is large and localized between training atoms (SI G.2). Even if $\hat{\boldsymbol{s}} - s_t^\star = -\nabla(\hat{E}_t - E_t^\star)$ is an exact score gap, it should not be equated directly with $\hat{v}_t - v_t^\star$ without the appropriate drift conversion and marginal-score terms.

E Finite-ring setup and landscape evidence

Training and metrics. The ring training set contains 16 equally spaced points on a unit circle, with held-out points at the angular midpoints. For the landscape of this section, a direct-score MLP is trained from random initialization with the denoising-score objective. Memorization on the ring is measured by an endpoint-distance threshold around the training points, not by the GMM/CelebA gap ratio, and the midpoint occupancy f_{mid} is the fraction of endpoints inside neighborhoods of the held-out midpoints. f_{mid} is not $1 - f_{\mathrm{mem}}$, because an endpoint may be near neither a training point nor a midpoint.

Projected potentials. At a fixed checkpoint and noise level σ, write $x = re_r(\theta)$ with $e_r = (\cos\theta, \sin\theta)$ and $e_\theta = (-\sin\theta, \cos\theta)$, let $\hat{s}_c = \hat{\boldsymbol{s}} \cdot e_c$ for $c \in \{r, \theta\}$, and let $\overline{\hat{s}}_c(r)$ be its angular average. The learned ring radius $\hat{r}$ satisfies $\overline{\hat{s}}_r(\hat{r}) = 0$ and $\partial_r \overline{\hat{s}}_r(\hat{r}) < 0$, and

$$E_{\mathrm{rad}}(r) = -\int_{\hat{r}}^{r} \overline{\hat{s}}_r(r')\,\mathrm{d}r', \qquad E_{\mathrm{ang}}(\theta) = -\hat{r}\int_{0}^{\theta} \left[\hat{s}_\theta(\hat{r}, \theta') - \overline{\hat{s}}_\theta(\hat{r})\right]\mathrm{d}\theta'. \tag{95}$$

Because the angular average is subtracted, E_{ang} is 2π-periodic. For an exact score $\hat{\boldsymbol{s}} = \nabla \log \hat{p}$ it equals $-\log \hat{p}(\hat{r}, \theta) + \log \hat{p}(\hat{r}, 0)$; for a general network field it is only a one-dimensional effective potential, not a global energy on the plane.

Broad well and narrow minima. Figure 6 and Table 1 show the landscape at $\sigma = 0.05$. After 10^4 updates the network has learned the ring (radius 1.010, radial depth 15.7) with an almost flat angular potential and little copying ($f_{\mathrm{mem}} = 0.070$). With further training the radial well deepens only moderately (15.7 to 18.9), while angular minima of depth 3.76 form at the training points and f_{mem} rises to 0.935. The broad structure (the ring) thus coexists with narrow training-point minima inside it, which is the situation assumed in Section 2.

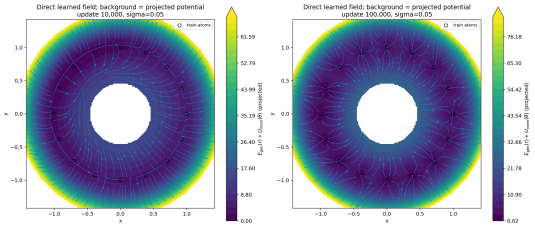


Figure 6: Score streamlines and projected potentials of the ring score network at 10^4 and 10^5 updates.

Table 1: Ring checkpoints at $\sigma = 0.05$: learned radius, radial well depth, angular potential difference between a training point and the adjacent midpoint, and copy fraction.

Training update	learned radius	radial depth	point–midpoint difference	f_{mem}
10k	1.010	15.724	0.0126	0.070
50k	1.004	18.312	2.902	0.871
100k	1.002	18.901	3.762	0.935

Sampling comparison. The comparison of Figure 2 uses a second ring model trained by flow matching, so that KTS can be applied in its original setting. All four methods use the same frozen checkpoint, training points and held-out midpoints. Churn changes the stochasticity, KTS rescales the velocity along its own direction, and HCI adds the tangential field of SI F.3. The baseline gives $f_{\mathrm{mem}} \approx 0.97$ and $f_{\mathrm{mid}} \approx 0.01$; at matched $f_{\mathrm{mem}} \approx 0.34$–$0.36$, the midpoint fractions are approximately 0.12 (churn), 0.05 (KTS) and 0.42 (HCI).

F HCI IMPLEMENTATION DETAILS

This section specifies the sampler of Section 4.4 for each experiment. The backbone is frozen in every case; HCI and the transport correction act only on its output.

subsectionNoise-prediction wrapper

The network predicts $\boldsymbol{\varepsilon}(x, t) = -\sigma_t \hat{\boldsymbol{s}}(x, t)$. Using the fitted reference score s_t^{ref}, we construct $u_t = -K s_t^{\mathrm{ref}}$. When the transport correction is applied, we use

$$\hat{\boldsymbol{s}}^{\,c} = \hat{\boldsymbol{s}} + n_t n_t^{\top}\left(s_t^{\mathrm{ref}} - \hat{\boldsymbol{s}}\right), \qquad n_t = \frac{s_t^{\mathrm{ref}}}{\max(\|s_t^{\mathrm{ref}}\|, 10^{-12})}. \tag{96}$$

The tangential score component is unchanged.

For an additive HCI update, the noise prediction becomes

$$\begin{aligned}\widetilde{\varepsilon} &= \varepsilon + n_t n_t^\top \big(\varepsilon^{\mathrm{ref}} - \varepsilon\big) - \eta w(t)\sigma_t u_t,\\ \varepsilon^{\mathrm{ref}} &= -\sigma_t s_t^{\mathrm{ref}}.\end{aligned} \tag{97}$$

The second term is omitted outside the transport interval, while $w(t)$ restricts HCI to the low-noise window. The intervals and additional scaling used in each experiment are specified below.

The modified prediction corresponds to the score $\hat{s}^{\,c} + \eta w(t) u_t$. In the decreasing-$t$ convention of (14), the HCI contribution to the drift coefficient is $-\eta g_t^2 w(t) u_t$ for the SDE and $-\eta g_t^2 w(t) u_t/2$ for the probability-flow ODE. These signs become positive when time increases in the sampling direction. The sampler otherwise follows its original update. With $\eta = 0$ and the transport correction switched off, the original noise prediction is recovered.

F.1 Fitted references and closed-form scores

For a Gaussian mixture $p^{\mathrm{ref}} = \sum_k \pi_k \mathcal{N}(m_k, \Sigma_k)$, the VP-noised path is again a Gaussian mixture with means $\sqrt{\bar\alpha_t} m_k$ and covariances $C_{k,t} = \bar\alpha_t \Sigma_k + (1-\bar\alpha_t) I$, so its score is (15) with responsibilities

$$\gamma_{k,t}(x) = \frac{\pi_k \, \mathcal{N}\big(x; \sqrt{\bar\alpha_t} m_k, C_{k,t}\big)}{\sum_j \pi_j \, \mathcal{N}\big(x; \sqrt{\bar\alpha_t} m_j, C_{j,t}\big)}. \tag{98}$$

For a single Gaussian $\mathcal{N}(m, \Sigma)$ with $\Sigma = V \operatorname{diag}(\lambda) V^\top$, $s_t^{\mathrm{ref}}(x) = -V \operatorname{diag}\big(1/(\bar\alpha_t \lambda_i + 1 - \bar\alpha_t)\big) V^\top (x - \sqrt{\bar\alpha_t} m)$. Both are exact gradient fields, so $\nabla s_t^{\mathrm{ref}}$ is symmetric and Proposition 3 applies with p^{ref} in place of $p^\star$. The normal is $n_t = s_t^{\mathrm{ref}} / \max(\|s_t^{\mathrm{ref}}\|, 10^{-12})$.

F.2 Constructed fields and admissible rescalings

The field is $u_t = -K s_t^{\mathrm{ref}}$ with a constant skew matrix K: $K_{\mathrm{block}} = \operatorname{diag}(K_2, K_2, K_2, K_2)$ on the GMM, and $K = \sum_{k=1}^{8} \big(v_{2k-1} v_{2k}^\top - v_{2k} v_{2k-1}^\top\big)$ on CelebA, with v_i the eigenvectors of the reference covariance in decreasing order. Any skew K gives $\nabla \cdot (p_t^{\mathrm{ref}} u_t) = 0$ identically (Proposition 3). The field must be neither normalized nor gated per sample: for a scalar function g, $\nabla \cdot (p_t^{\mathrm{ref}} g\, u_t) = p_t^{\mathrm{ref}} u_t \cdot \nabla g$, which vanishes only if g is constant along the flow of u_t. Only factors constant in x are therefore admissible: the time weight $w(t)$, the strength η, and on CelebA one scalar per step and batch that fixes the mean injection norm.

F.3 Ring field

For a rotationally symmetric reference, $u_t^{\mathrm{ring}}(x) := w(t) K_2 x / \|x\|$ with $K_2 = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ is tangential and divergence-free away from the origin, so $\nabla \cdot (p_t^\star u_t^{\mathrm{ring}}) = 0$ there; the ring endpoints lie far from the origin. It need not be orthogonal to $\hat{s}$ at a memorized checkpoint. The weight $w(t)$ is nonzero only at low noise levels, and the drift ηu_t^{ring} is added directly to the frozen flow-matching velocity, $\mathrm{d}x/\mathrm{d}t = \hat{v}(x,t) + \eta u_t^{\mathrm{ring}}(x)$; exact KTS is applied to the same velocity.

F.4 GMM: corrector integration

On the GMM the field is applied as a corrector. At each reverse step t, before the ancestral predictor, the state is transported by the flow $\mathrm{d}x/\mathrm{d}s = u_t(x)$ at fixed t over $s \in [0, \eta\, w(t) \beta_t]$. Since the reverse SDE has $g_t^2\, \mathrm{d}t = \beta_t$ per step, this is the added drift $\eta\, w(t)\, u_t$ of Theorem 1. The weight $w(t)$ increases quadratically from zero over the final 30% of the schedule. The flow is integrated with classical RK4, using $\lceil h L_t / 0.2 \rceil$ substeps for a step of length h, where $L_t = 1 + \bar\alpha_t \mu^2 d$. The network is re-evaluated after the corrector, and the normal correction (96) is applied inside the predictor for $t/T < 0.3$.

F.5 CelebA: additive injection

On CelebA the field enters the DDIM-100 noise prediction additively,

$$\tilde{\varepsilon} = \varepsilon + \eta\, w(t)\, c_t\, \sigma_t K s_t^{\mathrm{ref}}, \tag{99}$$

which is (97) with $u_t = -Ks_t^{\text{ref}}$. The weight w ramps linearly from 1 at $t = 0$ to 0 at $t = 0.25T$, and c_t is one scalar per step and batch that sets the mean norm of the injection to the mean norm of ε. The normal correction is applied at every reverse step, as $\tilde{\varepsilon} \leftarrow \tilde{\varepsilon} + n_t n_t^\top (\varepsilon^{\text{ref}} - \tilde{\varepsilon})$ with $n_t = \varepsilon^{\text{ref}}/\|\varepsilon^{\text{ref}}\|$ and $\varepsilon^{\text{ref}} = -\sigma_t s_t^{\text{ref}}$. It has no free parameter.

G GMM EXPERIMENTS

G.1 HCI WITH A REFERENCE FITTED TO THE TRAINING DATA

Setup. The target is $p^\star = \frac{1}{2}\mathcal{N}(\mu\mathbf{1}, I) + \frac{1}{2}\mathcal{N}(-\mu\mathbf{1}, I)$ in $\mathbb{R}^8$ with $\mu = 1$; its VP-noised score $s_t^\star(x) = -x + m_t \tanh(m_t \cdot x)$ with $m_t = \sqrt{\bar{\alpha}_t}\,\mu\mathbf{1}$ is used only for evaluation. The training set contains 256 points drawn from $p^\star$ with seed 1, the score network has width 128, and the checkpoint is evaluated at 4×10^6 updates. Every setting uses 4,000 samples of the ancestral sampler with 1,000 steps and the same initial and per-step noise (seed 0). KS distances of $\log p^\star$ are computed against 4,000 independent target draws (seed 777); two independent target sets differ by KS 0.034, which is the sampling floor. The covariance ratio is the per-cluster covariance trace, with clusters assigned by the sign of $\mathbf{1} \cdot x$, divided by its target value $d = 8$.

References. The two-component reference is a full-covariance Gaussian mixture fitted by expectation–maximization (200 iterations, initialized by splitting the training points by the sign of their projection on the leading principal direction, covariance regularized by $10^{-4}I$). It recovers the modes at axis coordinates $+2.88$ and -2.79 (true ± 2.83) with weights 0.501 and 0.499. The one-Gaussian reference is the sample mean and covariance of all training points: it is centered between the modes (axis coordinate $+0.05$), with variance 9.2 along the mode axis and about 1 in the orthogonal directions, so it is wrong along one direction only.

Protocol. HCI uses $u_t = -K_{\text{block}} s_t^{\text{ref}}$ with the corrector of SI F.4, and the correction (96) is applied for $t/T < 0.3$. The capped variant multiplies the correction by $\min(1, \rho\|s_t^{\text{ref}}\|/\|c\|)$, where c is the correction vector. "Reference samples" are drawn from the fitted reference itself. Our VP adaptation of KTS multiplies the deterministic displacement of each ancestral reverse step by $\zeta(s) = 1 - \beta_0\big(e^{3(s-0.6)} - 1\big)$, clamped positive, for sampling progress $s = 1 - t/(T-1) \geq 0.6$ ($s = 0$ at noise, $s = 1$ at data), and by $\zeta = 1$ before; the added noise is unchanged. This is the soft-landing schedule of KTS with its default split 0.6 and rate 3, without its early boost. Isotropic noise adds $\sigma_{\text{add}}\mathcal{N}(0, I)$ to the final baseline samples.

Table 2: HCI with a reference fitted to the training data, and comparison settings, on the memorized checkpoint with paired noise. KS and MMD^2 are against independent target draws (sampling floor: KS 0.034).

Setting	f_{mem}	KS	cov. ratio	MMD^2	d_1/ℓ_{train}
Baseline	.990	.034	1.03	2.1×10^{-3}	0.05
Reference samples, two-component	.000	.042	1.05	1.6×10^{-3}	–
HCI only, two-component	.013	.836	0.30	2.0×10^{-2}	0.61
HCI + capped correction ($\rho = 1$), two-component	.010	.796	0.33	1.8×10^{-2}	0.63
Normal correction only, two-component	.664	.031	1.04	2.1×10^{-3}	0.28
HCI + normal correction, two-component, $\eta = 300$	.025	.039	1.05	1.5×10^{-3}	0.82
HCI + normal correction, two-component, $\eta = 10^3$	.004	.033	1.03	1.5×10^{-3}	0.92
HCI + normal correction, two-component, $\eta = 3\times10^3$	.002	.050	1.05	1.3×10^{-3}	0.97
Reference samples, one Gaussian	.000	.217	1.34	4.9×10^{-3}	–
HCI only, one Gaussian	.030	.421	0.22	1.3×10^{-1}	0.61
HCI + capped correction ($\rho = 1$), one Gaussian	.025	.405	0.28	1.0×10^{-1}	0.62
Normal correction only, one Gaussian	.690	.027	1.03	2.1×10^{-3}	0.27
HCI + normal correction, one Gaussian, $\eta = 10^3$	.005	.128	1.14	5.6×10^{-3}	0.97
KTS (VP adaptation), $\beta_0 = 0.45$	.028	.078	1.07	4.7×10^{-3}	0.76
Noise, $\sigma_{\text{add}} = 0.525$	.019	.199	1.31	3.9×10^{-3}	0.78

Findings. The capped correction behaves like HCI alone. Along n_t, the network error at low noise exceeds $\|s_t^{\text{ref}}\|$ by far (SI G.2), so a cap at $\rho\|s_t^{\text{ref}}\|$ leaves most of it in place. The uncapped

correction works even with the misspecified reference, because the reference is much less wrong than the network along its own normal: on target draws at $t = 5$, the root-mean-square normal errors are 0.35 (one-Gaussian reference) and 309 (network); at $t = 100$ they are 0.31 and 6.8. Moreover, only 2.3% of the one-Gaussian normal lies along the mode axis, the one direction in which that reference is wrong. The two-mode structure itself is decided at high noise, where the correction is off; the remaining error of the one-Gaussian reference shows in the final samples as a bridge between the modes (KS 0.13).

G.2 Why HCI alone contracts the log-density marginal

By Proposition 3, the flow of a constructed field keeps every sample on its level set of the reference, so an HCI corrector step can redistribute probability within level sets, which is how it removes copies, but not across them. The distribution across level sets changes only in the predictor, through the component of the network score along the level-set normal. We measure this component with the radial pull $R(x) = s(x, t) \cdot (x - c)$, where c is the nearer noised cluster mean; its reference expectation under $p_t^\star$ is $-d = -8$. We compare points drawn from $p_t^\star$, which generically lie between training atoms, with noised training atoms, where baseline samples concentrate (4,000 points each).

t	σ_t	$\mathbb{E}_{p_t^\star}[\hat{R}]$	$\mathbb{E}_{p_t^\star}[R^\star]$	$\mathbb{E}_{\text{atoms}}[\hat{R}]$	$\mathbb{E}_{\text{atoms}}[R^\star]$
5	0.030	-846.0	-8.00	-1.8	-8.45
20	0.079	-264.0	-7.93	-7.7	-8.38
50	0.173	-76.4	-8.10	-7.4	-8.47
100	0.324	-25.0	-7.93	-7.8	-8.39
200	0.586	-10.0	-7.90	-7.8	-8.17

On the atoms the learned radial pull matches the reference, which is why the baseline, whose samples sit on atoms that are themselves target draws, has a nearly correct log-density marginal (KS 0.034). Between atoms the memorized score points toward the nearest atom, along a chord and therefore inward, and at low noise this pull exceeds the reference by up to two orders of magnitude. HCI keeps samples moving between atoms during the final 30% of the schedule, so they are exposed to this inward pull and the marginal contracts ("HCI only" in Table 2). In the notation of SI D.4, this is a large transport error localized between training atoms, which HCI does not correct by construction; the normal correction removes it.

Reproduction. All settings run on one frozen checkpoint (SHA-256 `dcce5f5a...3688`) in about three minutes on one RTX 4090. `gmm_fitted_ref_hci.py` and `gmm_fitted_ref_diag.py` produce Table 2, `plot_gmm_fitted_row.py` draws Figure 3, and `gmm_exact_hci_diag.py` computes the radial pull.

H CelebA experiments

H.1 Configurations, sampling and metrics

We follow the CelebA-32 protocol of Bonnaire et al. (2025): 32×32 grayscale images, a score U-Net of base width W trained from scratch on n images with the ordinary ε-prediction loss (SGD with momentum, batch 512, learning rate 0.01) to 2×10^6 updates. We use three memorized configurations: c08 ($n = 512$, $W = 32$; main text), c14 ($n = 512$, $W = 64$) and c07 ($n = 256$, $W = 32$), with baseline pixel copy fractions of 93.4%, 99.5% and 99.7%. Every setting draws 10,000 DDIM-100 samples with the seed convention of the benchmark, so all methods share the initial noise; bootstrap standard errors of f_{mem} are below 0.5 percentage points.

f_{mem} is the gap criterion $d_1/d_2 < 1/3$ in pixel space; std. f_{mem} applies it after normalizing every generated and training image to zero mean and unit variance. The NN ratio is the median distance to the nearest training image divided by the distance to the nearest of 512 held-out images (indices $100{,}000$–$100{,}511$ of the image tensor, disjoint from all training subsets), after the same normalization. Brightness and contrast KS distances compare the per-image mean and standard deviation of the samples with those of held-out indices $100{,}000$–$109{,}999$. FID is the mean over the five reference statistics of the benchmark.

Table 3: Full-run values of the c08 settings shown in Figure 4 (10,000 samples each).

Setting	FID	f_{mem} (%)	std. f_{mem} (%)	NN ratio	bright. KS	contr. KS
No intervention	18.8	93.4	93.5	0.11	0.03	0.04
HCI, fitted reference + normal corr., $\eta = 0.7$	40.3	4.6	**3.8**	0.54	0.03	0.16
CPSample, $\alpha = 0.49$, $s = 100$	42.5	28.4	26.9	0.43	0.07	0.27
Churn, $S_{\mathrm{churn}} = 41$	18.8	99.7	99.7	0.08	0.07	0.06
Held-out real faces	–	–	–	1.00	–	–

H.2 Fitted Gaussian reference

Early-network reference. For each configuration we take the latest checkpoint up to 1.05×10^5 updates whose baseline copy fraction is below 5% on the dense checkpoint grid: 1.05×10^5 for c08 ($f_{\mathrm{mem}} = 0.14\%$, FID 52.4), 8.5×10^4 for c14 and 7×10^4 for c07. From 10,000 DDIM-100 samples of that checkpoint we compute the mean m and covariance Σ in the network's input normalization and use the closed-form score of SI F.1. With 10,000 samples in $d = 1024$ the covariance has full rank. Sampling the Gaussian alone gives FID 147 on c08, so the reference is not itself a usable face model; sampling the early checkpoint alone gives FID 51–64.

Training-image reference. The alternative uses the n training images with Ledoit–Wolf shrinkage toward $(\mathrm{tr}\,\Sigma/d)I$, since $n < d$ (shrinkage 0.012 for $n = 512$ and 0.026 for $n = 256$).

Field. K rotates within the planes of eigenvector pairs $(v_1, v_2), \ldots, (v_{15}, v_{16})$ of Σ (SI F.2); injection and correction follow SI F.5.

H.3 Replication runs with a rule fixed in advance

The HCI design (PCA planes, rank 8, early-network Gaussian reference, window $0.25T$) was chosen on c08 with seed 0, the run of the main text, from the search listed in Table 5. We then fixed it and ran c08 with seed 1 and c14 and c07 with seed 0. Only the reference source and $\eta \in \{0.4, 0.5, 0.6, 0.7\}$ were varied for HCI; the comparison methods received widened grids that also vary their settings other than strength. Before these runs we wrote down the grids, the selection rule (for each method, the setting with the lowest std. f_{mem} among those with FID at most Δ above the baseline, $\Delta \in \{10, 25\}$) and the success criterion (HCI with the early-network reference is lowest in at least two of the three runs for both budgets); the verdict was computed mechanically from the saved metrics.

The criterion is met (Table 4); the declared comparison set also contained two earlier field constructions of ours that are not part of this paper, and the verdict is the same with or without them. HCI has the lowest value in every cell; in c07 at $\Delta = 10$ only its weakest setting is eligible, and neither it nor any other method removes copies there (92.2% versus 93.9% for CPSample). On c08 the replication reproduces the main-text run (4.3% versus 3.8% std. f_{mem} at $\eta = 0.7$), and the brightness KS of HCI stays within 0.01 of the baseline in every run. The training-image reference also removes copies without a brightness shift (std. f_{mem} 0.9%, 3.8% and 0.6% on c08, c14 and c07 at $\eta = 0.7$), but its FID is 19–52 points higher than with the early-network reference at the same η, so within the budgets it is eligible only on c08 and c14 at $\Delta = 25$.

H.4 Comparison methods

All methods share the checkpoint, the DDIM-100 grid, the seed convention and the evaluation code. KTS and churn samples are generated with the benchmark scripts. Our VP adaptation of KTS scales the deterministic displacement of each reverse step toward the posterior mean, as on the GMM (SI G.1). CPSample follows the authors' algorithm (confidence gate α, log-sigmoid stabilizer, optional time window); its random-label classifier was trained with the registered configuration and stopping criterion (15,000 updates; low-noise BCE 0.011 and 0.004, confident fraction 0.996 and 1.0 for the 512- and 256-image subsets).

Table 4: Std. f_{mem} (%) of the setting selected for each method under an FID budget Δ above the baseline, with its FID in parentheses; bold marks the lowest value per column, and a dash means that no setting met the budget.

Method	c08		c14		c07	
FID budget above baseline	+10	+25	+10	+25	+10	+25
No intervention	93.0 (18.8)		99.6 (20.8)		99.7 (30.0)	
HCI, fitted ref. (early network) + corr.	**24.0** (26.6)	**4.3** (40.5)	**15.3** (26.6)	**15.3** (26.6)	**92.2** (37.8)	**9.6** (50.0)
HCI, fitted ref. (training images) + corr.	–	49.5 (42.0)	–	48.2 (44.7)	–	–
CPSample	79.4 (24.0)	26.8 (42.8)	70.0 (30.1)	70.0 (30.1)	93.9 (34.8)	27.8 (52.3)
KTS (VP adaptation)	92.2 (25.8)	90.6 (40.3)	99.5 (28.4)	99.3 (37.7)	99.6 (35.4)	99.4 (53.8)
Churn	97.5 (18.4)	97.5 (18.4)	99.9 (20.5)	99.9 (20.5)	99.9 (30.2)	99.9 (30.2)

Table 5: Settings evaluated per configuration in the replication runs, and the search of the c08 run from which the HCI setting was taken.

Method	Settings per run	Count
HCI, fitted reference (ours)	reference $\in$ {early network, training images} $\times$ $\eta \in \{0.4, 0.5, 0.6, 0.7\}$; basis (PCA planes), rank 8, $\rho = 1$, window $0.25T$ fixed from the c08 selection run	8
c08 selection run (seed 0)	basis {PCA, DCT, random}, rank 2–128, reference {Gaussian, early network}, $\eta \in [0.3, 1.4]$	34
CPSample	$\alpha \in \{0.1, 0.25, 0.49\} \times s \in \{30, 100, 300\}$; time window $[0, 0.5T] \times \alpha = 0.49 \times s \in \{100, 300, 1000\}$	12
KTS (VP adaptation), $k = 3$	$\alpha_0 \in \{0, 0.01\} \times \beta_0 \in \{0.02, 0.05, 0.1, 0.15, 0.2\}$ ($\tau_s = 0.6$); $\tau_s \in \{0.4, 0.8\} \times \beta_0 \in \{0.05, 0.1, 0.15\}$ ($\alpha_0 = 0$)	16
Churn	$S_{\mathrm{churn}} \in \{6.4, 25.6, 41\} \times \{(S_{\mathrm{noise}} = 1$, all $\sigma)$, $(S_{\mathrm{noise}} = 1.007$, all $\sigma)$, $(S_{\mathrm{noise}} = 1, \sigma \in [0.05, 1])\}$	9

I Checks of alternative explanations

Table 6: Alternative explanations for a lower copy fraction and the tests that address them.

Alternative	Test	Result
Samples leave the data distribution	Exact log $p^\star$ histogram on the GMM (Table 2)	HCI + correction stays at the KS sampling floor with $f_{\mathrm{mem}} = 0.004$; isotropic noise reaches $f_{\mathrm{mem}} = 0.019$ only at KS 0.20.
Global intensity shift	Std. f_{mem} and brightness KS (Tables 3 and 4)	HCI: std. f_{mem} 3.8% versus pixel 4.6%, brightness KS equal to the baseline.
Copying a different training example	Paired columns with own nearest neighbor (Figure 4)	Churn copies another image in every column; HCI columns have $r = 0.48$–0.86.
The correction alone does the work	Correction without HCI (Table 2)	$f_{\mathrm{mem}} = 0.66$–0.69.
The reference generates the samples	Sampling the reference alone (SI H.2)	Gaussian: FID 147; early network: FID 51–64.
Selection on the reported samples	Replication with a rule fixed in advance (SI H.3)	Criterion met for both budgets; HCI lowest in all six cells.
Comparison methods undertuned	Widened grids (Table 5)	The extra settings did not change the best method in any cell; the HCI design search remains wider.

J Reproducibility, statistical scope, and compute

Evaluation sizes and seeds. GMM: 4,000 samples per setting with shared noise (seed 0) and 4,000 independent target draws for KS (seed 777; floor 0.034). CelebA: 10,000 DDIM-100 samples per setting with the benchmark seed convention; the main-text run uses seed 0, and the replication runs use seed 1 (c08) and seed 0 (c14, c07). Bootstrap standard errors of the CelebA copy fractions are below 0.5 percentage points; seed-to-seed variation was measured only on c08 (two seeds). The ring comparison uses one frozen flow-matching checkpoint for all methods.

Thresholds and selection. $\gamma_{\mathrm{mem}} = 1/3$ is inherited from the benchmark and was never fit on samples generated with HCI. The CelebA HCI setting of the main text was selected on the run that

also reports it; the replication runs used the rule and criterion of SI H.3, written down before they were run.

Compute. All numbers are computed on frozen checkpoints. The GMM experiments run in about three minutes on one RTX 4090. The CelebA runs use RTX 4090 GPUs with PyTorch 2.5.1 (CUDA 12.1) and deterministic cuDNN kernels. The reference is fitted once per configuration; at sampling time HCI and the correction add one closed-form score evaluation, a projection and a product with a rank-16 skew matrix per step, which is small compared with the U-Net evaluation.

Claim scope. The experiments concern near-exact copying of whole examples under the stated nearest-neighbor criterion. They do not prove differential privacy, rule out semantic identity leakage, or establish performance on large text-to-image systems. The constructed fields preserve the fitted reference exactly, not the unknown data distribution.