SCORE ADDITION IS NOT COMPOSITION: CLOSED-FORM PATH CORRECTIONS FOR TRAINING-FREE COMPOSITION OF PRETRAINED FLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained flow-matching models are becoming a dominant class of generative models, and there is a growing need to compose two of them without retraining, as a product, a geometric mean, or a negation. Current practice relies on heuristics, such as adding velocities or annealing along summed scores, or on particle-based correctors. The heuristics do not target the composition: adding velocities overshoots the product mean of two unit-variance Gaussians by a factor of $1+\pi/2$, and summed-score annealing follows the wrong intermediate curve, potentially misallocating mass across modes even with longer equilibration. We start from an exact identity. For flows trained with independent couplings from a common source, the intermediate marginals of any tilted composition factorize into powered component marginals and a kernel-corrected overlap integral of the endpoint posteriors, and the gradient of the log-overlap term is the exact correction to summed-score annealing. Gaussian moment matching gives the term a closed form that can be evaluated from black-box flows using only velocities and velocity Jacobians, and a well-posedness certificate identifies where the correction applies. Under frozen covariances the correction is a precision-weighted force that pulls samples toward agreement between the two models' endpoint predictions. The resulting sampler, CompFM, corrects the annealing path per sample and needs no particle population. At matched score-call budgets on class-intersection tasks with independently trained audit classifiers, CompFM attains 100.0% and 96.6% audited target-class purity on MNIST and FashionMNIST, respectively, versus 79.8% and 79.3% for summed-score annealing. On CIFAR-10, it concentrates 75.3% of audited mass on vehicle groups, compared with 1.1% for summed scores at the matched score-call budget. Composing two conditional flows on colored MNIST gives 96–99% joint accuracy, against 78–87% for summed scores. Classifier-free guidance (CFG) is the equal-covariance special case of the fusion rule, which also yields a covariance-aware correction to CFG.

1 INTRODUCTION

Modern generative modeling is consolidating around flow matching (Lipman et al., 2023; Liu et al., 2023; Albergo et al., 2025; Tong et al., 2024): industrial text-to-image systems (Esser et al., 2024) are trained as rectified flows with independent couplings from a Gaussian source. As pretrained flows accumulate, composition becomes a basic operation: given a flow for p^A and a flow for p^B, sample the product $p^A p^B$ (intersection of concepts, conjunction of constraints), a tempered geometric mean, or a contrastive negation, without retraining either model.

For diffusion models the usual answer is to add the scores, since $\nabla \log(p^A p^B) = \nabla \log p^A + \nabla \log p^B$ at the *endpoint*. At intermediate noise levels, however, the sum of the noised component scores generally differs from the score of the noised target product (Du et al., 2023), and recent work addresses this mismatch stochastically, with Feynman-Kac reweighting over particle populations (Skreta et al., 2025a) or Itô-density superposition (Skreta et al., 2025b). For flows, directly adding or averaging component velocities, v^A+v^B or $\frac{1}{2}(v^A+v^B)$, likewise provides no general guarantee of transporting the common source to the desired composition (Fig. 1).

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.

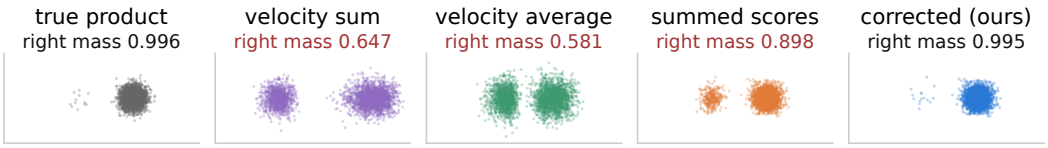


Figure 1: Composing an imbalanced bimodal flow (weights $0.65/0.35$) with an offset broad Gaussian. The true product moves almost all mass to the right mode; velocity sum, velocity average, and 480-call summed-score annealing all misallocate, while the corrected path matches the truth at the same budget (§6.1). All panels share axes, and oracle moments isolate the method from network error.

This paper supplies the structure that these heuristics lack. Our contributions are as follows.

1. **Exact identities (§3).** For flows trained with independent couplings from a common source, the endpoint posterior of any tilted composition is a kernel-corrected product of the individual posteriors (Thm. 1), and the intermediate marginals factorize as $p_t^{\alpha\beta} \propto (p_t^A)^\alpha (p_t^B)^\beta R_t$ (Thm. 2). The term $\nabla \log R_t$ is the exact error of summed-score annealing.
2. **A closed-form correction (§4).** Under Gaussian moment matching of the endpoint posteriors, whose means and covariances are available from any black-box flow via $m = x + (1-t)v$ and a velocity-Jacobian identity (Xing et al., 2026), the correction has a closed-form surrogate $\hat{R}_t$. We estimate its gradient by the frozen-covariance $\hat{g}_t$, which acts as a precision-weighted force toward agreement between the two models' endpoint predictions. A well-posedness certificate (Prop. 1) derived from Brascamp–Lieb identifies where the Gaussian fusion is well posed, and masking the correction outside this region keeps the method stable in practice.
3. **Theory (§5).** The fused *field* is exact for Gaussian families. We give perturbation bounds on the induced velocity error and its endpoint effect. We show that classifier-free guidance (Ho & Salimans, 2021) is the equal-covariance special case of our fusion, with a covariance-aware correction in general. We also prove a negative result: velocity addition overshoots the Gaussian product mean by $1 + \pi/2$ and doubles its variance.
4. **Evidence (§6).** On exactly solvable low-dimensional problems and on class-intersection tasks over MNIST, FashionMNIST, and CIFAR-10 audited by independently trained classifiers, the corrected path outperforms the heuristic baselines on every discriminating benchmark at matched budgets, including negation, geometric-mean, and conditional-conjunction tasks on real data. On learned models the summed-score path is non-monotone in its equilibration budget: more Langevin steps per slice can make the allocation worse, and longer annealing can drift through the reference, while the corrected path is stable across the deployed budget range.

The identities require independent couplings and a common source (Assumption 1), which is the rectified-flow recipe of current industrial models; scope and limitations are collected in §7.

2 Related Work

Flow matching and rectified flows. Simulation-free training of continuous flows (Lipman et al., 2023; Liu et al., 2023; Albergo et al., 2025; Tong et al., 2024) regresses a velocity field along stochastic interpolants; with independent couplings and a Gaussian source the score and the velocity are interconvertible (Lipman et al., 2024). Our work uses only black-box access to v_θ.

Composing diffusion models. Composable diffusion methods use score addition to combine pre-trained models (Liu et al., 2022); subsequent work showed that the summed scores generally fail to track the noised target product away from the endpoint (Du et al., 2023). SuperDiff (Skreta et al., 2025b) composes diffusions at inference with an Itô density estimator (logical OR/AND), and Feynman-Kac correctors (FKC) (Skreta et al., 2025a) sample products and annealed targets by reweighting particle populations on the *uncorrected* path. Our approach is complementary: we

correct the path itself, in closed form and per sample, and FKC-style weights can be layered on top (App. C). Training-time compositional approaches (Morshed & Boddeti, 2026) learn factorized velocity structure and are orthogonal.

Composition in simulation-based inference. The posterior for n i.i.d. observations is a product of single-observation posteriors tilted by a negative power of the prior, and its noised score is the summed scores plus an intractable term. F-NPSE (Geffner et al., 2023) anneals the summed scores; Linhart et al. (2026) close the term by fusing Gaussian denoising posteriors with score-Jacobian covariances, assuming the fused precision is positive definite. That construction reuses one amortized network with exponents fixed by n. Ours composes two independently trained flows with free weights, its shared factor is the path kernel rather than the prior, and the free weights are what carry the operator family: negation, geometric means, and covariance-aware guidance (§5) have no fixed-exponent counterpart. Prop. 1 moreover derives the positive-definiteness condition from log-concavity and tests it at every step.

Posterior-moment guidance. Approximations of denoiser posteriors are used for inverse-problem guidance to approximate the likelihood $p(y|x_t) = \int p(y|x_1)\, p(x_1|x_t)\, dx_1$ (Chung et al., 2023; Song et al., 2023; Finzi et al., 2023). Boys et al. (2024) study Jacobian-based posterior covariance estimates, whereas Bao et al. (2022) derive analytic reverse variances from score information. These guidance methods combine one model's posterior with a known likelihood or constraint, whereas we combine the posteriors of two learned models with a negative power of the shared kernel (Thm. 1); the flow covariance identity we use is due to Xing et al. (2026).

3 The exact structure of flow composition

Setup. A flow-matching model with linear path and $\sigma{=}0$ transports $x_0 \sim q_0$ to $x_1 \sim p_1$ along $x_t = (1{-}t)x_0 + tx_1$, with population-optimal velocity $v(t,x) = \mathbb{E}[x_1 - x_0 \mid x_t{=}x]$.

Assumption 1 (shared training configuration)**.** Models A, B are trained with the linear $\sigma{=}0$ path, independent coupling $q_0 \otimes p^i$, and common source $q_0 = \mathcal{N}(0, I_d)$.

This is the independent-coupling rectified-flow recipe; $\mathbb{E}_i$ denotes expectation under model i's joint law. We compose targets $p^{\alpha\beta} \propto (p^A)^\alpha (p^B)^\beta$ with $0 < Z_{\alpha\beta} = \int (p^A)^\alpha (p^B)^\beta < \infty$; $\alpha{=}\beta{=}1$ is the product, $\alpha{+}\beta{=}1$ the geometric family (containing CFG), and $\beta < 0$ negation. All statements refer to the *virtual model*: the flow-matching model with target $p^{\alpha\beta}$ under Assumption 1, whose exact velocity $v^{\alpha\beta}$ transports q_0 to $p^{\alpha\beta}$. Unless explicitly stated otherwise, the identities below are stated for $0 < t < 1$.

Lemma 1 (endpoint representation; any coupling)**.** $m^i(x,t) := \mathbb{E}_i[x_1|x_t{=}x] = x + (1{-}t)\, v^i(t,x)$ *and* $\mathbb{E}_i[x_0|x_t{=}x] = x - t\, v^i(t,x)$.

Lemma 2 (Bayes kernel; needs Assumption 1)**.** *With* $k_t(x|x_1) = \mathcal{N}(x; tx_1, (1{-}t)^2 I)$, $p^i(x_1|x_t{=}x) = p^i(x_1) k_t(x|x_1)/p_t^i(x)$*, and, as a function of* x_1*, the kernel is proportional to a Gaussian with mean* x/t *and precision* $\Lambda_k(t) = \frac{t^2}{(1-t)^2} I$.

Lemma 3 (score and covariance identities)**.** $s_t^i(x) = \frac{t\, v^i(t,x) - x}{1-t}$ *and* $\Sigma^i(x,t) = \mathrm{Cov}_i[x_1|x_t{=}x] = \frac{(1-t)^2}{t}\big[I + (1{-}t)\nabla_x v^i(t,x)\big]$.

The covariance identity is due to Xing et al. (2026); we re-derive it in App. A and use it in a new role.

Remark 1 (general Gaussian paths; diffusion models)**.** The structural identities below hold verbatim for *any* Gaussian path kernel $k_t(x|x_1) = \mathcal{N}(x; \alpha_t x_1, \sigma_t^2 I)$ upon setting $\Lambda_k = \frac{\alpha_t^2}{\sigma_t^2} I$, hence for pretrained diffusion models under the (α_t, σ_t) reparameterization (App. A.6); we instantiate the linear $\sigma{=}0$ flow case throughout.

The following two statements are exact.

Theorem 1 (posterior fusion; exact)**.** *Under Assumption 1, for all* (x,t) *with positive densities,*

$$p^{\alpha\beta}(x_1 \mid x_t{=}x) \;\propto\; \big[p^A(x_1|x)\big]^\alpha \big[p^B(x_1|x)\big]^\beta \, k_t(x|x_1)^{\,1-\alpha-\beta}. \tag{1}$$

Theorem 2 (marginal factorization and the exact score correction)**.** *Let* $R_t(x) = \int [p^A(x_1|x)]^\alpha [p^B(x_1|x)]^\beta k_t(x|x_1)^{1-\alpha-\beta} dx_1$. *Then*

$$p_t^{\alpha\beta} \propto (p_t^A)^\alpha (p_t^B)^\beta R_t, \qquad s_t^{\alpha\beta} = \alpha s_t^A + \beta s_t^B + \nabla \log R_t. \tag{2}$$

Equation (2) identifies $\nabla \log R_t$ as the error of summed-score annealing, not an approximation of it. At $t{=}0$, $\nabla \log R_0 = (1{-}\alpha{-}\beta)\nabla \log q_0$ cancels the double counting of the source ($\alpha s_0^A + \beta s_0^B = (\alpha{+}\beta)\nabla \log q_0$, while the correct score is $\nabla \log q_0$). At $t \to 1$ the correction vanishes for smooth positive targets. The correction therefore acts at intermediate t (Fig. 7; Fig. 6, App. A).

4 CompFM: a closed-form path correction

R_t involves the unknown posteriors. Our approximation is Gaussian moment matching: replace $p^i(x_1|x)$ by $\mathcal{N}(m^i, \Sigma^i)$ with the moments of Lemmas 1–3. Under this Gaussian approximation, the fused posterior and the overlap surrogate $\hat{R}_t$ have closed forms whenever $\Lambda^{\alpha\beta} \succ 0$:

$$\Lambda^{\alpha\beta} = \alpha\Lambda^A + \beta\Lambda^B + (1{-}\alpha{-}\beta)\Lambda_k, \quad m^{\alpha\beta} = (\Lambda^{\alpha\beta})^{-1}\big[\alpha\Lambda^A m^A + \beta\Lambda^B m^B + (1{-}\alpha{-}\beta)\Lambda_k \tfrac{x}{t}\big], \tag{3}$$

with $\Lambda^i = (\Sigma^i)^{-1}$, and $\log \hat{R}_t$ the explicit quadratic form (5) (App. A).

The deployed estimator. Differentiating $\log \hat{R}_t$ exactly would require $\nabla_x \Sigma^i \propto \nabla_x^2 v^i$. We therefore deploy the *frozen-covariance* gradient $\hat{g}_t$: Σ^i enters under stop-gradient (on images, refreshed once per slice) and the gradient flows only through the posterior means and the explicit kernel terms. This estimator requires velocities and velocity Jacobians alone, and $\hat{g}_t = \nabla \log \hat{R}_t$ whenever Σ is x-independent; the oracle experiments use the full gradient. In high dimension we additionally use the scalar (trace) covariance $\Sigma^i \approx \frac{(1-t)^2}{t}\big(1 + (1{-}t)\frac{\nabla \cdot v^i}{d}\big) I$ with a Hutchinson divergence probe. Table 5 isolates both approximations (§6.4).

What the correction does. With frozen covariances the surrogate has a pairwise form (App. A.8). For the product with scalar precisions λ_A, λ_B and kernel precision $\lambda_k = t^2/(1{-}t)^2$, it reads, up to a term that does not depend on x,

$$\log \hat{R}_t(x) = -\frac{1}{2\bar{\Lambda}}\Big[\lambda_A\lambda_B\|m^A{-}m^B\|^2 - \lambda_A\lambda_k\|m^A{-}\tfrac{x}{t}\|^2 - \lambda_B\lambda_k\|m^B{-}\tfrac{x}{t}\|^2\Big], \qquad \bar{\Lambda} = \lambda_A{+}\lambda_B{-}\lambda_k, \tag{4}$$

Hence $\hat{g}_t$ is a precision-weighted force that pulls x toward agreement between the two models' endpoint predictions, minus a term that removes the double-counted source. For the geometric family only the agreement term remains, and for classifier-free guidance its coefficient changes sign, so the correction becomes a repulsion between the conditional and the unconditional prediction. Class intersection requires this agreement. Along the anneal, the corrected path lowers $\|m^A{-}m^B\|^2$ at mid-t by $6.7\times$ on MNIST and by about $2\times$ on M2 relative to the summed-score path (Fig. 7).

Proposition 1 (well-posedness certificate)**.** *The fusion* (3) *is well posed iff* $\Lambda^{\alpha\beta} \succ 0$. *(i) The geometric family* ($\alpha, \beta \geq 0$, $\alpha{+}\beta{=}1$) *is always well posed. (ii) For the product, log-concavity of* one *factor suffices, by Brascamp–Lieb:* $\Sigma^A \preceq \Lambda_k^{-1}$, *hence* $\Lambda^A{+}\Lambda^B{-}\Lambda_k \succeq \Lambda^B \succ 0$. *(iii) For the product family, failure at* x *means that the posteriors are jointly more diffuse than the kernel along some direction* ($v^\top(\Lambda^A{+}\Lambda^B)v \leq v^\top\Lambda_k v$)*; empirically this is a mid-*t *mode-competition region (§6). Negative-weight operators can additionally fail through the negative precision coefficient, even for concentrated posteriors (Table 1).*

In the sampler the certificate is evaluated at every step for the precisions that are actually fused, scalar in high dimension, and the correction is switched off where it fails, falling back to the plain score.

The CompFM sampler. We anneal Langevin dynamics with the drift $\hat{s}_t^{\alpha\beta} = \alpha s_t^A + \beta s_t^B + \hat{g}_t$, which is the score of the corrected curve $\gamma_t \propto (p_t^A)^\alpha (p_t^B)^\beta \hat{R}_t$ when Σ is x-independent and an estimate of it otherwise, with the mask of Prop. 1 and a trust-region clip (both guards are ablated in §6.4; Algorithm 1 in App. B lists the deployed steps). We finish with a deterministic *fused-velocity tail* $\dot{x} = (m^{\alpha\beta}{-}x)/(1{-}t)$ on a late interval (e.g. $t \in [0.95, 0.995]$; App. B), where the moment-matched

field is provably accurate (§5); the tail removes residual Langevin noise. The sampler is population-free. An optional Feynman-Kac sequential Monte Carlo (SMC) layer over γ_t adds approximate annealed-importance reweighting (App. C) and is compatible with FKC (Skreta et al., 2025a).

Remark 2 (score-conversion amplification)**.** If $\|v_\theta - v\| \le \varepsilon$, then the converted score error is at most $\frac{t\varepsilon}{1-t}$ (Lemma 3): score conversion multiplies the velocity error by $\frac{t}{1-t}$. This explains the late-t divergence of score-based composition on learned flows, motivates ending the anneal early, and makes the velocity-based tail the appropriate finisher.

5 THEORY

Theorem 3 (exactness and error bound; informal)**.** *If p^A, p^B are Gaussian, the moment-matched fused field is exact for every operator in the family. In general, if the true posteriors are δ_t-close in log-density to their moment-matched Gaussians and $\Lambda^{\alpha\beta} \succeq \lambda_t I$ with $\lambda_t > 0$, the fused velocity error is at most $(e^{2(|\alpha|+|\beta|)\delta_t} - 1)\sqrt{d/\lambda_t}/(1-t)$ pointwise. Given a uniform spatial Lipschitz bound on the fused field and time-integrability of this envelope, the endpoint Wasserstein error obeys a Grönwall bound. Moreover, if $p^A, p^B \in C^4$ with locally bounded log-derivatives (nonnegative weights $\alpha, \beta \ge 0$), the moment-matched fused mean agrees with the true fused mean to $O\big((1-t)^4\big)$, so the fused velocity error decays as $O\big((1-t)^3\big)$ as $t \to 1$ (Lemma 5). The certificate of Prop. 1 tests positive-definiteness and does not locate approximation error.*

Corollary 1 (CFG is the equal-covariance degenerate case)**.** *For $\alpha = 1+\gamma$, $\beta = -\gamma$ (kernel-free geometric family), if $\Sigma^A = \Sigma^B$ then* (3) *reduces to $v = (1+\gamma)v^{\mathrm{cond}} - \gamma v^{\mathrm{uncond}}$, i.e., standard classifier-free guidance (guidance strength γ, distinct from the curve γ_t). With unequal covariances CompFM yields a covariance-aware CFG. The precision extrapolation $(1+\gamma)\Lambda^A - \gamma\Lambda^B$ can lose positive-definiteness at large γ; we evaluate covariance-aware CFG on the solvable family.*

Proposition 2 (velocity addition provably fails)**.** *For source $\mathcal{N}(0,1)$ and targets $\mathcal{N}(\mu_A, 1), \mathcal{N}(\mu_B, 1)$ the product is $\mathcal{N}(\frac{\mu_A+\mu_B}{2}, \frac{1}{2})$, while the ODE with v^A+v^B terminates at $\mathcal{N}\big((\frac{1}{2}+\frac{\pi}{4})(\mu_A+\mu_B),\, 1\big)$. The mean is overshot by $1+\pi/2 \approx 2.571$ and the variance is doubled; the total-variation gap is at least 0.16 even when $\mu_A+\mu_B=0$. Velocity* averaging *instead targets the geometric mean $\sqrt{p^A p^B}$ under equal covariances, which explains its partial success.*

All proofs are in App. A. Every closed-form constant above is verified numerically in §6; for example, the measured naive mean factor is 1.272 against 1.285 predicted, and the 1% gap is consistent with time truncation.

6 EXPERIMENTS

Setup. Benchmarks: exactly solvable compositions with oracle moments (A3 and M4 in 1D; M1 and M2 in 2D) and their learned-flow versions; class intersection on MNIST, FashionMNIST, and CIFAR-10; the operator family and conditional conjunction on MNIST. Baselines: velocity addition and averaging (ODE), summed-score annealing (the plain path), and FKC-style SMC. Metrics: W_2 and allocation error against exact targets on the oracle suite; audited purity on images, with auditors trained independently of all flows on the raw training sets. Unless stated otherwise, budgets are matched in *score-curve calls*, written NFE (one call evaluates every factor model once; correction overhead is measured in §6.4), and wall-clock is reported separately. Reported standard deviations are over sampling seeds at fixed checkpoints.

6.1 EXACTLY SOLVABLE COMPOSITION

Figure 1 shows the discriminating benchmark (M2). Naive velocity composition misallocates severely (right-mass $0.647/0.581$ vs. true 0.996), summed-score annealing at 480 NFE reaches 0.898, and the corrected path reaches 0.995. Across budgets (Fig. 2a) the corrected path's allocation error stays well below the plain path's ($7\times$ at 48 score calls, $\sim 80\times$ at 480), and at 48 calls it is already below the plain path at 480. On the symmetric benchmark M1 the correction has little effect on allocation error, since symmetry fixes the target left/right mass split (5-seed stds in Table 3, App. B).

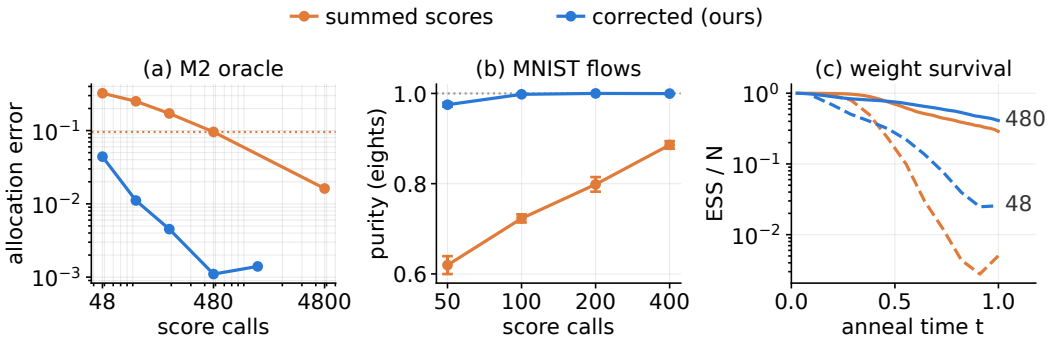


Figure 2: Budget scaling and weight survival, unweighted sampling. (a) M2 oracle allocation error against the score-call budget; the corrected path at 48 calls beats the summed-score path at 480 (dotted line). (b) MNIST purity against the slice-axis budget. (c) Normalized effective sample size (ESS/N) along the anneal without resampling, at 48 (dashed) and 480 (solid) calls. Orange is summed scores, blue the corrected path.

Table 1: W_2 distance to the exact tilted target for four operators, using ODE sampling with oracle velocities (n=4096). Lower is better.

operator	CompFM fusion	velocity lin. comb.	score lin. comb.
product	**0.076**	0.743	0.853
geometric 0.5/0.5	**0.108**	0.302	0.302
geometric 0.7/0.3	**0.112**	0.258	0.258
negation 1/−0.3	**0.173**	1.446	3.713

Learned flows. Replacing oracle moments by 3-layer MLPs trained with I-CFM (App. B; single-model W_2 floors 0.36/0.17) preserves the ordering on M2. The corrected anneal reaches right-mass 0.9934±0.0009 at 480 NFE (true 0.9957; $W_2 = 0.32\pm0.04$), beating the plain path at ten times the budget (0.9808±0.0035; $W_2 = 0.57$). On A3 the corrected sampler moves little across budgets (0.9903±0.0008 → 0.9980), while longer plain annealing drifts through the learned-product reference (0.9719 → 0.9995 as the budget grows 10×). This is the learned-level signature of the non-monotonicity of §6.3 (Table 4, App. B).

Adversarial stress test. M4 consists of two widely separated tri-modal factors on offset grids, where the Prop. 1 certificate disables the correction over 63% of the density-weighted mass at t=0.2. The corrected sampler still keeps a 3× advantage in per-mode allocation TV at matched budget (0.035±0.006 vs. 0.109±0.008 at 480 NFE, 5 seeds). It reaches 0.0065±0.002 at 1200 NFE, where the plain path needs 4800 NFE for 0.019±0.004, and without any weights it beats plain-path SMC with N=256 interacting particles (0.066±0.011).

6.2 Operator family and covariance-aware CFG

On exactly solvable anisotropic Gaussian pairs (cond. $\mathcal{N}((1, 0.5), \mathrm{diag}(0.5, 1))$ vs. uncond. $\mathcal{N}(0, 2I)$), standard CFG has endpoint mean error growing from 0.146 (γ=0.5) to 0.400 (γ=3), while the covariance-aware fusion stays at the sampling floor (0.048 → 0.036; Fig. 6 right, App. A). On a fixed Gaussian pair, the fusion is at the sampling floor for the product, geometric means, and negation, where naive velocity and score combinations err by up to $W_2 = 3.7$ (Table 1). The same operators on the MNIST flows of §6.3 (Table 6) give the intended targets: negation $p^A(p^B)^{-0.3}$ removes B's classes from A (97.2±0.8% threes with 0.2% eights, against 78.8% for summed scores and 84.8% for the velocity combination, which leave 11–12% eights), and the geometric mean reaches 99.3±0.5% eights where velocity averaging, its equal-covariance special case, reaches 53.2%.

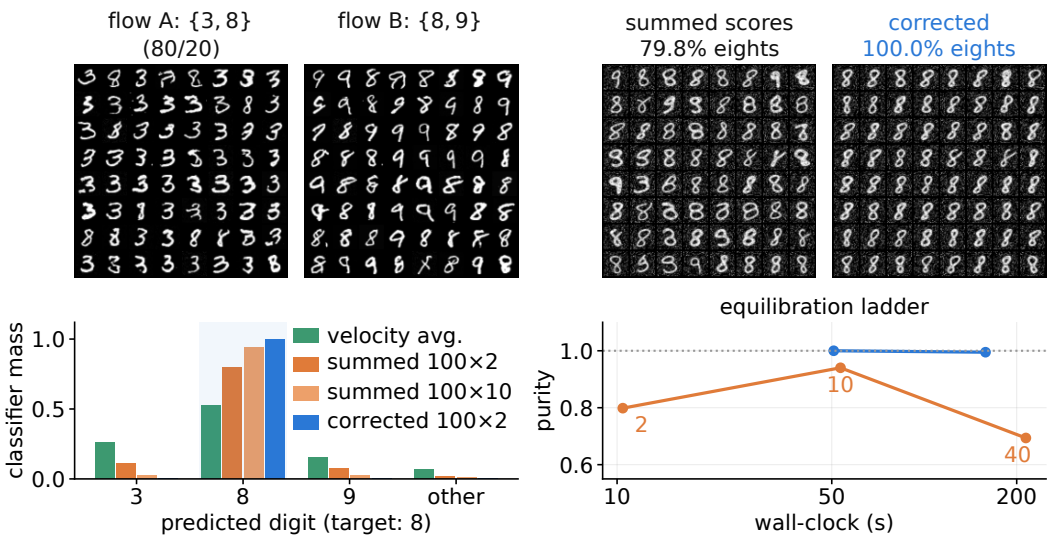


Figure 3: MNIST intersection. Flow A is imbalanced $80/20$ towards threes and flow B covers eights and nines, so the product should be pure eights. Top: the corrected sampler (blue) audits as pure eights where summed scores leak threes. Bottom left: the class-resolved audit, target class shaded; velocity averaging spreads mass over both parents' private classes. Bottom right: the plain path is non-monotone in equilibration steps at fixed 100 slices, while the corrected path stays near ceiling at $5\times$ compute. Grids: seed 0.

Table 2: Audited purity (%) on the image tasks at matched score-call budgets (slices $\times$ steps; 5 seeds). Velocity averaging is a 200-step ODE, run on MNIST and colored MNIST. CIFAR-10 at the 200-call budget: 60.7 ± 2.2 vs. 3.0 ± 0.5.

task (audited class)	corrected	summed scores	velocity avg.	budget
MNIST $\{3,8\}\times\{8,9\}$ (eights)	$\mathbf{100.0\pm0.0}$	79.8 ± 1.6	52.4	100×2
FashionMNIST (boots)	$\mathbf{96.6\pm0.5}$	79.3 ± 0.7	–	100×2 + tail
CIFAR-10 (vehicles)	$\mathbf{75.3\pm1.3}$	1.1 ± 0.4	–	200×2 + tail
colored MNIST, $8\times$red (both)	$\mathbf{96.2\pm0.8}$	87.1 ± 1.5	38.5 ± 2.7	100×2
colored MNIST, $3\times$green (both)	$\mathbf{98.5\pm0.7}$	77.8 ± 1.5	42.2 ± 1.2	100×2

6.3 Class-intersection on real data with independent audits

MNIST. Flow A is trained on $\{3,8\}$ with an $80/20$ imbalance and flow B on $\{8,9\}$, so the product must be pure eights. A CNN audit (98.7% test accuracy) gives, over 5 seeds: velocity averaging 52.4% eights; summed-score annealing $79.8\pm1.6\%$ (equal steps), $94.0\pm1.1\%$ (equal wall-clock), and $69.4\pm2.3\%$ at $4\times$ wall-clock, so the equilibration axis is non-monotone. Scaling the slice axis instead improves the plain path monotonically ($62.0\pm2.0\%$ at 50 calls to $88.6\pm0.8\%$ at 400, Fig. 2b), yet leaves it 11 points below CompFM at half that budget (CompFM passes $97.5\pm0.4\%$ already at 50 calls). CompFM reaches 100.0% at equal steps (Table 2) and $99.5\pm0.3\%$ at $5\times$ compute (Fig. 3), with diversity preserved: corrected samples are narrower than the class marginal, as a product should be, with no coverage collapse (99.6% nearest-neighbor coverage of real eights; App. B). A step-size-scaled control ties the excursion to total equilibration time rather than step count. Refining the step at fixed time leaves purity unchanged, while at the 40-step point's time it deepens the collapse to $37.9\pm3.9\%$ (App. B). This is consistent with amplified late-t score error (Rem. 2) acting on the biased slice targets; the corrected path is stable under the identical schedule through $5\times$.

FashionMNIST. Sneaker/boot $\times$ boot/sandal $\to$ boots; both methods use the same fused-velocity denoising tail (§4). CompFM reaches $96.6\pm0.5\%$ boots against $79.3\pm0.7\%$ for the plain path (Table 2), with diverse silhouettes (Fig. 4).

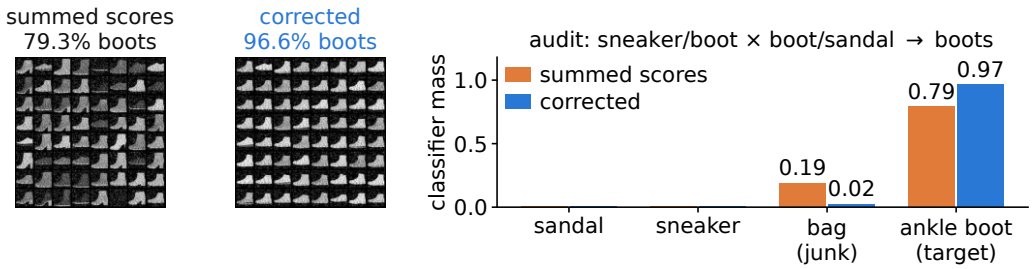


Figure 4: FashionMNIST intersection, both methods with the same fused-velocity tail. The corrected path gives varied ankle-boot silhouettes; the plain path leaves substantial mass on noise artifacts that the auditor reads as bags. Grids: seed 0.

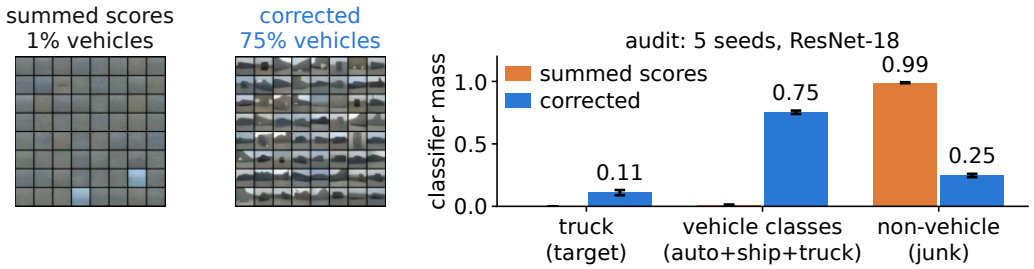


Figure 5: CIFAR-10 intersection. Left: the plain summed-score path composes two individually well-behaved flows (App. B) into featureless gray fields. Middle: the corrected path yields coherent vehicle scenes. Right: audited mass over the target, vehicle, and junk groups (5 seeds). Grids: seed 0.

Non-monotonicity across datasets. The same compute ladder (5 seeds) shows the same instability on every dataset. On FashionMNIST the plain path collapses from $79.3{\pm}0.7\%$ to $8.3{\pm}0.6\%$ at $5\times$ steps ($7.4{\pm}0.5\%$ at $20\times$). On CIFAR-10 it is erratic and never recovers ($3.0{\pm}0.5 \to 0.4{\pm}0.3 \to 7.0{\pm}1.3\%$ vehicle mass at $1/5/20\times$, all $\ll$ the corrected 60.7%). The corrected path is stable across the deployed range ($99.5{\pm}0.3\%$ MNIST, $96.7{\pm}0.4\%$ FashionMNIST at $5\times$). This is consistent with Rem. 2: longer equilibration increases exposure to the biased slice-wise score fields.

CIFAR-10. Automobile/truck $(80/20)$ $\times$ truck/ship $\to$ trucks; 64-channel UNet flows (50k steps, EMA), ResNet-18 audit (88.9% test). The plain path produces featureless fields that the audit scatters over animal classes (vehicle mass $3.0{\pm}0.5\%$, degrading to $1.1{\pm}0.4\%$ at double budget). The corrected path concentrates the recognizable mass on vehicle classes (Table 2; trucks $12.4{\pm}2.1$ / $11.0{\pm}2.1\%$ at 200/400 calls; Fig. 5). A folder-FID audit against real trucks corroborates the gap (FID 246 vs. 399; all three datasets in App. B). The corrected path thus recovers vehicle-group structure; within-group resolution is discussed in §7 and App. D.

Conditional conjunction. In practice one usually composes two conditionals of the same distribution. On MNIST with strokes tinted red, green, or blue at random, we train a digit-conditional and a color-conditional flow and compose $p(x \mid \text{digit})\, p(x \mid \text{color})$. Audited jointly by a digit CNN and the dominant stroke channel (5 seeds, Tables 2 and 7), the corrected sampler reaches 96–99% joint accuracy where summed-score annealing reaches 78–87%, velocity averaging 39–42%, and the single conditionals 10–38%.

6.4 Ablations, cost, and Feynman-Kac correctors

Table 5 (App. B) crosses frozen vs. full gradients with scalar vs. full covariance on the oracle suite: freezing forfeits part of the gain there. All image results use the deployed frozen-scalar estimator; no full-gradient comparison is available on images. Both guards are necessary. Removing the trust-region clip collapses MNIST purity to $77.9{\pm}1.1\%$, the plain level, and disabling the Prop. 1 mask crashes the oracle sampler wherever an ill-posed region exists. At matched score-call budgets, the

deployed image correction costs 4.7–$4.8\times$ per anneal. On MNIST, it still improves audited purity at equal wall-clock (100% vs. 94.0%).

Feynman-Kac correctors. On the M2 oracle the corrected path without any weights already matches plain-path SMC with $N{=}256$ particles, and on MNIST FKC needs a global 512-particle population and 100 slices to match our weight-free sampler; the two methods also compose (App. C).

7 LIMITATIONS

Our identities require independent couplings and a shared Gaussian source (Assumption 1); minibatch-OT-coupled flows break the kernel structure (App. E), and extending to OT couplings is open. The correction inherits network moment error, and the deployed frozen-covariance estimator forfeits part of the gain on low-dimensional oracles (App. B). Single-Gaussian matching cannot represent multimodal posteriors; the mask guards the fused-precision condition but does not certify approximation accuracy, and a mixture-aware extension is future work. Negation with the roles of A and B exchanged did not succeed in our MNIST runs.

Class-intersection audits are classifier-based (one auditor per dataset): purity certifies membership as judged by that auditor, not closeness to the learned product, since images admit no tractable reference (the learned-flow suite provides that test, and such metrics are standard (Du et al., 2023; Skreta et al., 2025a)). Our largest models are CIFAR-scale UNets, where the flows resolve the vehicle group but not the target class: most recognizable mass is audited as ship. An audit using model log-likelihood estimates (App. D) locates the cause in the learned densities rather than the sampler. Corrected outputs are statistically indistinguishable under the learned product across audited classes, which gives no evidence of a sampler-side class bias, while the pixel-space likelihoods themselves rank smooth off-class images above the target class. This inversion is unchanged under $4\times$ model capacity and $2\times$ training duration, consistent with known complexity biases of pixel-space likelihoods (Nalisnick et al., 2019; Kirichenko et al., 2020). The limitation therefore concerns pixel-space density products as composition targets, which any faithful sampler shares, rather than the correction; we read CIFAR as recovering class-space structure rather than a completed intersection, and latent-space composition is a candidate direction.

The optional weight layer has limited scope: on the adversarial M4, under heavy masking, it is $4\times$ worse than the weight-free sampler (App. C), so we recommend the weight-free sampler in certificate-poor regimes. The reported image experiments are at most $32{\times}32$; at higher resolution the anneal protocol and the covariance estimates would need to be revisited, and directional (diagonal or low-rank) covariance estimation is the open direction. Finally, learned-score annealing is intrinsically fragile near $t{=}1$ (Rem. 2). At extreme over-equilibration ($20\times$ steps) even the corrected anneal degrades on FashionMNIST ($21.1{\pm}1.8\%$), hence the deployed short-anneal+tail protocol; the tail mitigates but does not remove the fragility.

8 CONCLUSION

Composition of pretrained flows has an exact structure: the intermediate marginals of any tilted composition factorize through an overlap term R_t. Gaussian moment matching yields a closed-form surrogate $\hat{R}_t$, whose frozen-covariance gradient estimator $\hat{g}_t$ corrects the annealing path. The fused field is exact on Gaussian families. The corrected path reaches near-ceiling audited purity on MNIST and FashionMNIST class-intersection tasks and supports conditional conjunction on MNIST, with budget advantages over summed-score annealing in the reported comparisons. The endpoint-moment view should extend to other training-free operations on pretrained flows.

AI USE STATEMENT

In this work, we used generative AI tools to help formulate mathematical claims, to assist in the writing of proofs, and to help implement part of the methods. We have not used generative AI tools to generate synthetic data sets, to help develop theoretical models or conceptual frameworks, to provide critical ingredients for proving mathematical claims, to propose or refine hypotheses, to design or provide feedback on research methodology or experiments, to assist with translation, to

clean and reformat datasets, to support qualitative and thematic data analysis, or to interpret results. We have carefully reviewed and checked all AI-assisted work.

Reproducibility statement

The standing assumption is Assumption 1, complete proofs of every theoretical claim are in App. A, and the full experimental protocol, including datasets, architectures, sampler constants, and seeding, is in App. B.

References

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

Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-DPM: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *International Conference on Learning Representations*, 2022.

Benjamin Boys, Mark Girolami, Jakiw Pidstrigach, Sebastian Reich, Alan Mosca, and Omer Deniz Akyildiz. Tweedie moment projected diffusions for inverse problems. *Transactions on Machine Learning Research*, 2024.

Herm Jan Brascamp and Elliott H. Lieb. On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22(4):366–389, 1976.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Pierre Del Moral. *Feynman-Kac formulae: genealogical and interacting particle systems with applications*. Springer, 2004.

Yilun Du, Conor Durkan, Robin Strudel, Joshua B. Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and MCMC. In *International Conference on Machine Learning*, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Marc Anton Finzi, Anudhyan Boral, Andrew Gordon Wilson, Fei Sha, and Leonardo Zepeda-Núñez. User-defined event sampling and uncertainty quantification in diffusion models for physical dynamical systems. In *International Conference on Machine Learning*, 2023.

Tomas Geffner, George Papamakarios, and Andriy Mnih. Compositional score modeling for simulation-based inference. In *International Conference on Machine Learning*, 2023.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2019.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

Polina Kirichenko, Pavel Izmailov, and Andrew G. Wilson. Why normalizing flows fail to detect out-of-distribution data. In *Advances in Neural Information Processing Systems*, 2020.

Julia Linhart, Gabriel Cardoso, Alexandre Gramfort, Sylvain Le Corff, and Pedro L. C. Rodrigues. Diffusion posterior sampling for simulation-based inference in tall data settings. *Transactions on Machine Learning Research*, 2026.

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.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, 2022.

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

Mashrur M. Morshed and Vishnu Boddeti. Compositional generative modeling from decentralized data. In *International Conference on Machine Learning*, 2026.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019.

Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *International Conference on Machine Learning*, 2023.

Marta Skreta, Tara Akhound-Sadegh, Viktor Ohanesian, Roberto Bondesan, Alan Aspuru-Guzik, Arnaud Doucet, Rob Brekelmans, Alexander Tong, and Kirill Neklyudov. Feynman-Kac correctors in diffusion: Annealing, guidance, and product of experts. In *International Conference on Machine Learning*, 2025a.

Marta Skreta, Lazar Atanackovic, Joey Bose, Alexander Tong, and Kirill Neklyudov. The superposition of diffusion models using the Itô density estimator. In *International Conference on Learning Representations*, 2025b.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

Jiarui Xing, Song Wang, and Jian Wang. Uncertainty in a single pass: A closed-form identity for one-step flow matching. *arXiv preprint arXiv:2605.00941*, 2026.

A Proofs

A.1 Lemma 1

With $\sigma{=}0$, $x = (1{-}t)\mathbb{E}[x_0|x_t{=}x] + t\,\mathbb{E}[x_1|x_t{=}x]$ and $v = \mathbb{E}[x_1|x_t] - \mathbb{E}[x_0|x_t]$ form a linear system in the two conditional expectations; solving it gives both identities. No coupling assumption is used. □

A.2 Lemma 2

Given x_1, $x_t = (1-t)x_0 + tx_1$ with $x_0 \sim \mathcal{N}(0, I)$ independent of x_1 (Assumption 1), so $x_t|x_1 \sim \mathcal{N}(tx_1, (1-t)^2 I)$ and Bayes' rule gives the posterior form. Completing the square, $\|x - tx_1\|^2 = t^2\|x_1 - x/t\|^2$, exhibits k_t as a Gaussian in x_1 with mean x/t and precision $\Lambda_k = \frac{t^2}{(1-t)^2} I$. Under a coupling with $\pi(x_0|x_1) \neq q_0$ the kernel depends on x_1 through π and the shared-kernel structure is lost. □

A.3 Lemma 3

Tweedie's formula for the Gaussian channel $x_t|x_1$ gives $s_t(x) = \frac{t\, m(x,t) - x}{(1-t)^2}$; substituting Lemma 1 yields $s_t = (tv - x)/(1-t)$. Differentiating the posterior in x, $\nabla_x p(x_1|x) = p(x_1|x)\big[\frac{tx_1 - x}{(1-t)^2} - s_t(x)\big]$, hence $\nabla_x m = \frac{t}{(1-t)^2}\Sigma$; combined with $\nabla_x m = I + (1-t)\nabla_x v$ this gives the covariance identity, which matches Xing et al. (2026). Sanity check (target $\mathcal{N}(\mu, 1)$, $s_t \equiv t^2 + (1-t)^2$): the identity returns $(1-t)^2/s_t$, equal to the Gaussian conditional variance $1 - t^2/s_t$. □

A.4 Theorem 1

By Lemma 2, $p^i(x_1) = p^i(x_1|x)\, p_t^i(x)/k_t(x|x_1)$. Substituting into $p^{\alpha\beta}(x_1|x) \propto (p^A(x_1))^\alpha (p^B(x_1))^\beta k_t(x|x_1)$ and absorbing the x_1-independent factor $(p_t^A)^\alpha (p_t^B)^\beta$ into the normalizer gives (1). □

A.5 Theorem 2

$p_t^{\alpha\beta}(x) = Z_{\alpha\beta}^{-1} \int (p^A)^\alpha (p^B)^\beta k_t \, dx_1$; expanding with the same substitution yields $(p_t^A)^\alpha (p_t^B)^\beta R_t$ and the score identity follows by taking $\nabla \log$. Limits: at $t{=}0$ the kernel is constant in x_1 and independence gives $p^i(x_1|x) = p^i(x_1)$, so $R_0(x) = q_0(x)^{1-\alpha-\beta} \cdot \text{const}$; the correction $(1{-}\alpha{-}\beta)\nabla \log q_0$ restores $s_0^{\alpha\beta} = \nabla \log q_0$ (verified numerically: fused score $\approx -x$, summed score $\approx -2x$). □

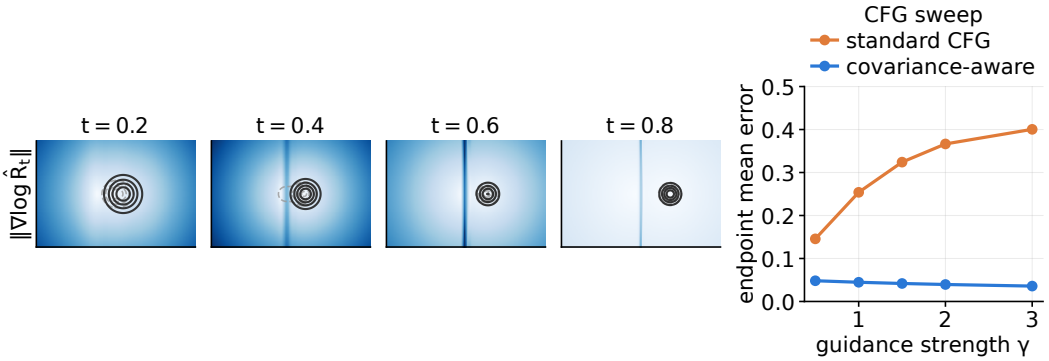


Figure 6: Mechanism. Left: the size of the correction, $\|\nabla \log \hat{R}_t\|$, over space at four times on M2; gray dashed contours are the plain-path slice, solid contours the true composed marginal. The correction concentrates along the mode-competition ridge and decays as the posteriors localize, and the certificate of Prop. 1 is positive everywhere on this benchmark. Right: guidance sweep on anisotropic Gaussians; standard CFG's endpoint mean error grows with the guidance strength, while the covariance-aware fusion stays at the sampling floor.

Remark 3 (more than two models)**.** All formulas extend linearly to K factors $\prod_i (p^i)^{\alpha_i}$: the kernel exponent in (1) becomes $1 - \sum_i \alpha_i$, and the fusion (3) sums K precision terms.

A.6 General Gaussian paths

For a kernel $k_t(x|x_1) = \mathcal{N}(x; \alpha_t x_1, \sigma_t^2 I)$ the kernel precision is $\Lambda_k = \frac{\alpha_t^2}{\sigma_t^2} I$, and for ε-prediction diffusion models one reads off $m = (x - \sigma_t \hat{\varepsilon})/\alpha_t$ and $\Sigma = \frac{\sigma_t^2}{\alpha_t^2}\big(I - \sigma_t \nabla_x \hat{\varepsilon}\big)$, so the identities, fusion rule, and certificate carry over verbatim. Deploying the full sampler there additionally involves the reversed time convention and the $\alpha_t \to 0$ endpoint, which we do not evaluate in this paper.

A.7 The explicit quadratic form of $\log \hat{R}_t$

Writing $c = x/t$, $\Lambda_k = \frac{t^2}{(1-t)^2} I$, $\kappa = 1 - \alpha - \beta$, $\Lambda^{\alpha\beta} = \alpha\Lambda^A + \beta\Lambda^B + \kappa\Lambda_k$ and $b = \alpha\Lambda^A m^A + \beta\Lambda^B m^B + \kappa\Lambda_k c$, Gaussian moment matching of the overlap integral gives, up to an additive function of t alone,

$$\begin{aligned}\log \hat{R}_t(x) = \tfrac{1}{2}\, b^\top (\Lambda^{\alpha\beta})^{-1} b - \tfrac{1}{2}\big[\alpha\, m^{A\top}\Lambda^A m^A + \beta\, m^{B\top}\Lambda^B m^B + \kappa\, c^\top \Lambda_k c\big] \\ - \tfrac{1}{2}\log\det\Lambda^{\alpha\beta} + \tfrac{1}{2}\big[\alpha \log\det\Lambda^A + \beta\log\det\Lambda^B\big],\end{aligned} \tag{5}$$

with every m^i, Λ^i evaluated at (x, t) via Lemma 3. The scalar variant substitutes $\Lambda^i = \lambda_i(x,t) I$ from the divergence probe. The deployed frozen-covariance gradient (§4) differentiates (5) treating Λ^A, Λ^B (hence both determinant terms) as constants in x, keeping the dependence through m^A, m^B and the explicit c-terms; the oracle gates differentiate (5) exactly. This is the formula the implementation evaluates verbatim for the product, and in scale-stable scalar form in high dimension.

A.8 Pairwise form of the frozen-covariance surrogate

For scalar weights $w_1, \ldots, w_K$ of any sign with $W = \sum_i w_i \neq 0$ and points p_i, $\frac{1}{2}\|\sum_i w_i p_i\|^2/W - \frac{1}{2}\sum_i w_i \|p_i\|^2 = -\frac{1}{2W}\sum_{i<j} w_i w_j \|p_i - p_j\|^2$. Applied to (5) with scalar precisions $\Lambda^i = \lambda_i I$ (kernel precision $\lambda_k = t^2/(1-t)^2$), weights $(\alpha\lambda_A, \beta\lambda_B, \kappa\lambda_k)$, and points $(m^A, m^B, x/t)$, the x-dependent part of $\log \hat{R}_t$ under frozen precisions is $-\frac{1}{2\bar{\Lambda}}\big[\alpha\beta\lambda_A\lambda_B\|m^A - m^B\|^2 + \alpha\kappa\lambda_A\lambda_k\|m^A - x/t\|^2 + \beta\kappa\lambda_B\lambda_k\|m^B - x/t\|^2\big]$ with $\bar{\Lambda} = \alpha\lambda_A + \beta\lambda_B + \kappa\lambda_k$. The product ($\kappa = -1$) gives (4); the geometric family ($\kappa = 0$) keeps only the agreement term; for CFG ($\alpha = 1+\gamma$, $\beta = -\gamma$) the coefficient $\alpha\beta$ is negative and the term is a repulsion. The identity agrees with the deployed implementation to 10^{-14}. Figure 7 tracks $\|m^A - m^B\|^2/d$ along the anneal under the canonical protocols (MNIST: 100 slices $\times$ 2 steps, 3 seeds; M2: exact moments, 120 slices $\times$ 4 steps, 3 seeds).

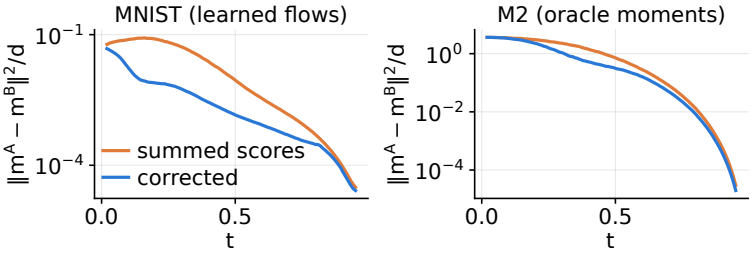


Figure 7: Disagreement between the two models' endpoint predictions along the anneal (mean over particles; band: one standard deviation over seeds). Left: MNIST learned flows. Right: the M2 oracle. The corrected path holds the two predictions 6.7× (MNIST) and 2× (M2) closer at mid-anneal, which is where the composition target is decided.

A.9 Proposition 1

(i) A convex combination of positive-definite matrices is positive definite. (ii) If p^A is log-concave its posterior is a log-concave prior times a Gaussian likelihood with precision Λ_k, so $-\nabla^2 \log p^A(x_1|x) \succeq \Lambda_k$; by Brascamp–Lieb (Brascamp & Lieb, 1976), $\Sigma^A \preceq \Lambda_k^{-1}$, hence

$\Lambda^A \succeq \Lambda_k$ and $\Lambda^A+\Lambda^B-\Lambda_k \succeq \Lambda^B \succ 0$ (Λ^B inverts a nondegenerate covariance). With both factors log-concave the stronger bound $\succeq \Lambda_k$ holds. At t=0, Λ_k=0 and $\Lambda^i = \mathrm{Cov}(p^i)^{-1} \succ 0$, provided the targets have finite second moments with full-rank covariance (a mild standing assumption; alternatively restrict to $t \in (0,1)$, which is all the sampler uses). (iii) is the contrapositive of the display in (3). Gaussian self-product check (A=B=$\mathcal{N}(\mu,\varsigma^2)$): the fusion returns precision $2\varsigma^{-2}+\Lambda_k$ and precision-weighted mean $2\mu/\varsigma^2+\Lambda_k x/t$, exactly the virtual-model posterior of the target $\mathcal{N}(\mu,\varsigma^2/2)$. □

A.10 Corollary 1

With $\alpha+\beta$=1 the kernel term vanishes; with Λ^A=Λ^B the fused mean is $(1+\gamma)m^A - \gamma m^B$, and $v = (m-x)/(1-t)$ is affine in m. □

A.11 Proposition 2

For target $\mathcal{N}(\mu,1)$ the oracle field is $v(t,x) = \frac{\mu(1-t)+x(2t-1)}{s_t}$, $s_t = t^2+(1-t)^2$ (its mean ODE has the exact solution $\bar{m}(t) = t\mu$ and variance s_t, confirming the machinery). For $\dot{x} = v^A+v^B$ the mean obeys $\dot{\bar{m}} = [\mu_\Sigma(1-t)+2\bar{m}(2t-1)]/s_t$ with integrating factor $1/s_t$ (since $\frac{d}{dt}\ln s_t = \frac{2(2t-1)}{s_t}$), so $\bar{m}(1) = \mu_\Sigma \int_0^1 \frac{1-t}{s_t^2}dt$. Splitting $1-t = \frac{1}{2}+(\frac{1}{2}-t)$: the second part integrates to $\frac{1}{4}[1/s_t]_0^1 = 0$; for the first, $\int_0^1 s_t^{-2}dt = 1+\frac{\pi}{2}$ by the standard $\int \frac{du}{(au^2+c)^2}$ formula with $s = 2u^2+\frac{1}{2}$, $u = t-\frac{1}{2}$. Hence $\bar{m}(1) = (\frac{1}{2}+\frac{\pi}{4})\mu_\Sigma$. The variance obeys $\dot{\Sigma} = \frac{4(2t-1)}{s_t}\Sigma$, so $\Sigma(1) = e^{2[\ln s_t]_0^1} = 1$. The TV bound is $\mathrm{TV}(\mathcal{N}(0,1),\mathcal{N}(0,\frac{1}{2})) \approx 0.166$ by direct computation. Velocity averaging is the α=β=$\frac{1}{2}$ equal-covariance case of (3), whose target is $\sqrt{p^A p^B}$; for the same pair this is $\mathcal{N}(\frac{\mu_A+\mu_B}{2},1)$, matching the measured endpoint $(0.494, 1.012)$. □

A.12 Perturbation lemma and Theorem 3

Lemma 4. *If $\pi \propto e^{-U}$, $\tilde{\pi} \propto e^{-U-\epsilon}$ with $\|\epsilon\|_\infty \le \delta$, then $d\tilde{\pi}/d\pi \in [e^{-2\delta}, e^{2\delta}]$ and $\|\mathbb{E}_{\tilde{\pi}}x - \mathbb{E}_\pi x\| \le (e^{2\delta}-1)\sqrt{\mathrm{tr}\,\Sigma_\pi}$.*

Proof. The normalizer ratio lies in $[e^{-\delta}, e^{\delta}]$, giving the density-ratio band. Since $\int(d\tilde{\pi} - d\pi) = 0$, $\mathbb{E}_{\tilde{\pi}}x - \mathbb{E}_\pi x = \mathbb{E}_\pi[(x-m_\pi)(\frac{d\tilde{\pi}}{d\pi}-1)]$; bound the second factor uniformly and apply Cauchy–Schwarz. □

For Theorem 3: Gaussian exactness is immediate, since all factors in (1) are Gaussian and moment matching is lossless. Explicitly, for targets $p^i = \mathcal{N}(\mu_i, \Lambda_i^{-1})$ the posterior precisions are $\Lambda^i = \Lambda_i + \Lambda_k$, so the fused precision is $\alpha\Lambda_A + \beta\Lambda_B + \Lambda_k$ and the fused precision-weighted mean is $\alpha\Lambda_A\mu_A + \beta\Lambda_B\mu_B + \Lambda_k x/t$, which are exactly the virtual-model posterior moments of the tilted target $\mathcal{N}\big((\alpha\Lambda_A+\beta\Lambda_B)^{-1}(\alpha\Lambda_A\mu_A+\beta\Lambda_B\mu_B), (\alpha\Lambda_A+\beta\Lambda_B)^{-1}\big)$; the fusion is well posed for every t because $\alpha\Lambda_A+\beta\Lambda_B \succ 0$ is precisely the normalizability of the tilted target (the standing condition $Z_{\alpha\beta} < \infty$ of §3). In general the fused log-density perturbation is at most $(|\alpha|+|\beta|)\delta_t$, and with $\Lambda^{\alpha\beta} \succeq \lambda_t I$, $\mathrm{tr}\,\Sigma \le d/\lambda_t$; dividing the mean error by $(1-t)$ gives the velocity bound. Coupling the exact and approximate ODEs from a common initial point, $\frac{d}{ds}\|\hat{x}-x\| \le L\|\hat{x}-x\| + \|\Delta v\|$ and Grönwall give the endpoint bound; taking L^2 expectations over the common source bounds W_2. □

Remark 4 (scope of Theorem 3)**.** The theorem controls the *field*, the moment-matched fused velocity, and not the sampler built on it (fixed-step unadjusted Langevin, mask, clip, frozen covariances, endpoint truncation), whose layers are ablated in §6.4. The sup-norm log-ratio hypothesis is restrictive for multimodal posteriors (tails differ), so on GMM benchmarks the bound is a perturbation principle where the log-ratio is controlled, not an unconditional guarantee; the oracle gates play that role empirically.

A.13 Late-time accuracy: the Laplace regime

Lemma 5 (cubic late-time decay)**.** *Fix α, β and write $\rho = (1-t)/t$, $c = x/t$. Suppose $\alpha, \beta \ge 0$, p^A, p^B are bounded on $\mathbb{R}^d$, C^4 on a ball $B_r(c)$ with $p^i(c) \ge p_{\min} > 0$ and $\|\nabla^k \log p^i\| \le L_k$ on*

$B_r(c)$ *for* $k \le 4$ *(for negative weights, additional global domination of* $(p^i)^\beta$ *is needed and we do not pursue it). Then, as* $\rho \to 0$*,*

$$\big\|\hat{m}^{\alpha\beta}(x,t) - m^{\alpha\beta}(x,t)\big\| = O(\rho^4), \qquad \big\|\hat{v}^{\alpha\beta}(x,t) - v^{\alpha\beta}(x,t)\big\| = O\big((1-t)^3\big),$$

where $\hat{m}^{\alpha\beta}$ *is the Gaussian fusion* (3) *evaluated at the exact posterior moments and* $m^{\alpha\beta}$ *is the exact fused mean.*

Proof. Write $g_i = \nabla \log p^i(c)$, $H_i = \nabla^2 \log p^i(c)$, $\bar{g} = \alpha g_A + \beta g_B$, $\bar{H} = \alpha H_A + \beta H_B$. *Step 1 (single-model moments).* By Lemma 2, $p^i(x_1|x) \propto p^i(x_1)\exp(-\|x_1 - c\|^2/2\rho^2)$. Substituting $x_1 = c + \rho u$ and expanding $\log p^i(c+\rho u) = \log p^i(c) + \rho\, g_i^\top u + \frac{\rho^2}{2} u^\top H_i u + R_i(u)$ with $|R_i(u)| \le \frac{L_3}{6}\rho^3\|u\|^3$ for $\|u\| \le r/\rho$, the law of u is proportional to $\exp(-\frac{1}{2}u^\top(I - \rho^2 H_i)u + \rho\, g_i^\top u + R_i(u))$; mass outside $B_r(c)$ is $O(e^{-r^2/8\rho^2})$ (bounding p^i above by its global supremum, finite by hypothesis, and below by $p_{\min}e^{-L_1 r}$ on the ball) and is absorbed into every displayed remainder. A linear tilt leaves a Gaussian's covariance unchanged, so the quadratic–linear part gives $\mathrm{Cov}(u) = (I - \rho^2 H_i)^{-1} = I + \rho^2 H_i + O(\rho^4)$ exactly. The cubic remainder perturbs $\mathrm{Cov}(u)$ only at $O(\rho^4)$: under C^4, $R_i(u) = \frac{\rho^3}{6}\nabla^3 \log p^i(c)[u,u,u] + O(\rho^4\|u\|^4)$, a *fixed* odd cubic plus a quartic-order remainder, so the cubic part pairs odd monomials with the symmetric base law (odd moments vanish) and its first surviving contribution is the cross term with the linear tilt, of order $\rho \cdot \rho^3$. The $O(\rho^4\|u\|^4)$ part and R_i^2 enter at $O(\rho^4)$ and $O(\rho^6)$ directly (all $\mathbb{E}\|u\|^m$ finite). Under C^3 alone the fixed-cubic split is unavailable and the same argument yields only $\Sigma_i = \rho^2(I + \rho^2 H_i) + O(\rho^5)$, i.e., a quadratic velocity rate; the C^4 assumption yields the cubic rate. Likewise $\mathbb{E}u = \rho g_i + O(\rho^3)$. Hence $m_i = c + \rho^2 g_i + O(\rho^4)$ and $\Sigma_i = \rho^2(I + \rho^2 H_i) + O(\rho^6)$, which the inversion in Step 3 requires. *Step 2 (exact fused mean).* The fused posterior $\propto (p^A)^\alpha (p^B)^\beta \exp(-\|x_1 - c\|^2/2\rho^2)$ admits the same expansion with $(g_i, H_i) \to (\bar{g}, \bar{H})$, hence $m^{\alpha\beta} = c + \rho^2 \bar{g} + O(\rho^4)$. *Step 3 (Gaussian fusion of exact moments).* From Step 1, $\Lambda_i = \Sigma_i^{-1} = \rho^{-2} I - H_i + O(\rho^2)$, hence $\Lambda^{\alpha\beta} = \rho^{-2} I - \bar{H} + O(\rho^2)$ (the ρ^{-2} coefficients sum to one across the three terms of (3)), and the fusion right-hand side is $\rho^{-2}c + \bar{g} - \bar{H}c + O(\rho^2)$. Multiplying by $(\rho^{-2} I - \bar{H})^{-1} = \rho^2(I + \rho^2 \bar{H}) + O(\rho^6)$,

$$\hat{m}^{\alpha\beta} = c + \rho^2 \bar{g} - \rho^2 \bar{H} c + \rho^2 \bar{H} c + O(\rho^4) = c + \rho^2 \bar{g} + O(\rho^4)\,:$$

the curvature terms cancel exactly, so the two means agree to $O(\rho^4)$. Dividing by $(1-t)$ gives the velocity rate, $\rho^4/(1-t) = (1-t)^3/t^4$. □

The cancellation in Step 3 is the reason the deterministic fused-velocity tail is the principled late-time finisher (§4); Figure 8 checks the rate on GMM benchmarks.

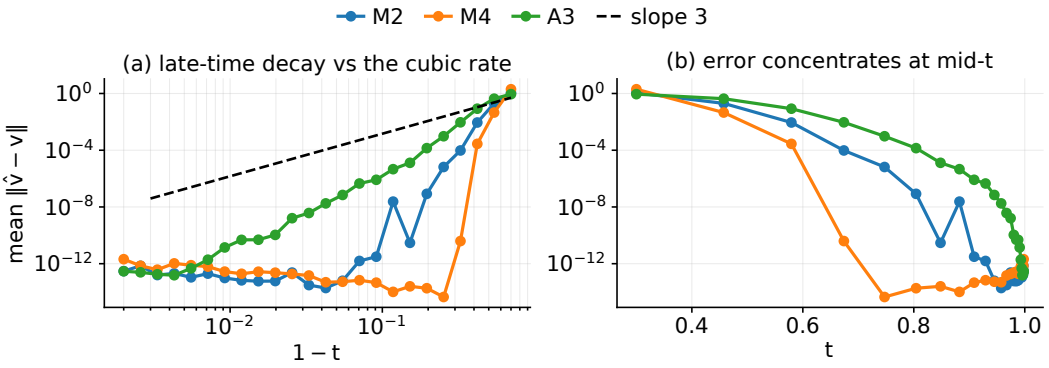


Figure 8: Empirical check of Lemma 5. Because a product of GMMs is a GMM, the true fused velocity is exactly computable; we plot the mean error of the moment-matched fusion against it along the virtual path. (a) Late-time decay is at least as fast as the guaranteed cubic rate (dashed guide); on these benchmarks the local quadraticity of the log-densities gives superconvergence down to machine precision. (b) The error concentrates at mid-t on these benchmarks.

B EXPERIMENTAL DETAILS

Table 3: Oracle suite (5 seeds; n=4096; exact posterior moments). The corrected anneal at 480 calls beats the plain anneal at 4800 on M2 and is neutral on the symmetric control M1. The score-sum-velocity conversion collapses variance; its low M2 allocation error reflects the allocation metric's insensitivity to within-mode variance collapse.

	A3 (1D, true 0.977)		M2 (2D, true 0.996)		M1 (symmetric)	
method	W_2	alloc. err	W_2	alloc. err	W_2	alloc. err
velocity sum	1.90±.02	.319±.006	3.57±.03	.353±.007	1.04±.01	.002±.001
velocity average	1.64±.02	.404±.006	2.64±.02	.421±.004	.306±.017	.002±.001
score-sum velocity	.775±.010	.023±.000	.936±.015	.004±.000	.880±.003	.002±.001
ULA plain @480	.481±.044	.057±.008	1.18±.03	.094±.003	.099±.006	.005±.004
ULA ours @480	.083±.031	**.005±.003**	.124±.050	**.001±.001**	.102±.014	.007±.002
ULA plain @4800	.068±.037	.002±.002	.482±.027	.018±.001	.095±.020	.007±.007
ULA ours @1200	.046±.025	.001±.001	.120±.046	.001±.001	.098±.024	.009±.006

Table 4: Learned flows (3-layer MLPs, I-CFM, 5 seeds). On A3 the appropriate reference is the *learned* product; the plain path at 480 NFE happens to sit near its right-mode mass and then overshoots it with more compute.

	M2 learned (true rm 0.9957)		A3 learned (learned ref 0.9729)	
method	W_2	right-mass	W_2	right-mass
ULA plain @480	1.18±.04	.9248±.0036	.310±.017	.9719±.0020
ULA ours @480	**.323±.043**	**.9934±.0009**	.387±.024	.9903±.0008
ULA plain @4800	.566±.060	.9808±.0035	.570±.016	.9995±.0003
ULA ours @1200	.350±.030	.9953±.0007	.528±.014	.9980±.0009

Low-dimensional suite. GMM oracles in float64; RK4 with 300 steps on $[10^{-2}, 1-10^{-3}]$ for velocity methods (the deterministic fused-velocity baseline drops the kernel term on certificate-failing rows, which restores positive-definiteness for the product family); annealed ULA with 120 slices $\times$ 4 steps (480 NFE) unless stated, step size $0.1\,(t^2c+(1-t)^2)$; the Prop. 1 mask disables the correction where the smallest eigenvalue of $\Lambda^A+\Lambda^B-\Lambda_k$ is below 0.1 (unscaled precision units; the same floor gates the SMC targets in App. C); W_2 by exact EMD (POT) on 4096 samples; 5 seeds. Learned flows: 3-layer MLPs (width 64), I-CFM with σ=0; learned velocity Jacobians are symmetrized and eigenvalue-floored before inversion (the population Jacobian is symmetric, a learned one need not be), and the finite-difference correction gradient costs $2d$ extra $\log \hat{R}_t$ evaluations per call. References are computed against the *learned* product (from 10^5 ODE samples per model with kernel density estimates). The A3 case shows why: the learned models are near-unbiased, yet long plain annealing overshoots both the analytic and the learned-product references. Retraining the MLPs shifts plain-ULA rows by up to 0.03 right-mass while corrected rows move ≤ 0.002, so the corrected-vs-plain conclusion is robust to the training draw; tabled values are from one such run. **Images.** MNIST/FashionMNIST: 32-channel UNets (Lipman et al., 2023)-style, 4k steps, Adam 10^{-3}; CIFAR-10: 64-channel UNet, 50k steps, EMA 0.9999, warmup 2500. Audits: MNIST CNN (98.7% test), Fashion CNN (89.7%), CIFAR ResNet-18 (88.9%). Sampler (Algorithm 1): 100 slices $\times$ 2 steps on $[0.02, 0.95]$, correction cutoff t=0.9, trust-region clip at the base-score norm, and the Prop. 1 mask applied to the scaled fused precision $\lambda_A+\lambda_B-t$ with floor 0.1 (scaled units, i.e., $0.1\,t/(1-t)^2$ in the unscaled units of the low-dimensional suite; it never binds on MNIST). MNIST runs anneal-only (no tail); FashionMNIST adds the fused-velocity tail on $[0.95, 0.995]$ (25 steps) to both methods; the tail floors the scaled fused precision at 0.1 as a numerical guard rather than re-applying the mask (late-t certificates are positive in practice). CIFAR-10: anneal on $[0.02, 0.90]$, correction cutoff t=0.85, tail on $[0.90, 0.999]$ (60 steps). CIFAR auditor on real test data: truck 92.7%, ship 94.0%, automobile 94.3% per-class accuracy; vehicle-group closure 96.9%. Score conversions cap t at 0.995 (images; 0.999 for the learned MLPs) below the endpoint singularity of Rem. 2. The deployed scalar algebra is pinned to the float64 full-covariance oracle by an exact consistency test: in the K=1 isotropic case, where the scalar form is exact and the Hutchinson divergence estimate has zero variance, the image-side implementation must match the oracle at both call

sites (score correction and fused tail) to relative error 10^{-8}. Wall-clock on a single H100. Diversity audits: mean pairwise L_2 and nearest-neighbor coverage vs. real intersection-class samples; FID, diversity, and coverage all use the main sampler protocol of their dataset (MNIST anneal-only; Fashion/CIFAR with tail), n=512, one sampling seed. Every image replicate is fully seeded: a stable per-(config, seed) value seeds the global torch RNG (Hutchinson probes, including the tail's) before the replicate, while the sampler's own generator drives initial and Langevin noise, so any config/seed pair reruns independently in a fresh process, and every log line records the derived seed. Verified: an independent fresh-process rerun of the Fashion main config reproduced all printed digits, and the Fig. 2a assets regenerate bitwise. NN coverage: for each of 512 real intersection-class images, we test whether its nearest generated sample (pixel-space L_2, n=512 generated) lies within twice its nearest *real* neighbor's distance; the $2r$ threshold is a heuristic choice and the criterion is pixel-space only.

Algorithm 1 Deployed CompFM sampler (image protocol; the product $\alpha{=}\beta{=}1$ is shown, general (α, β) replaces the weights as in App. A.8). The tail is used on FashionMNIST and CIFAR-10; MNIST runs anneal-only.

Require: flows v^A, v^B; slices $t_1 < \cdots < t_K$ in $[t_0, t_{\text{end}}]$; steps per slice S; step size $\eta(t) = \eta_c\,(c\,t^2 + (1{-}t)^2)$; cutoff t_{cut}; floor ρ; tail interval $[t_{\text{end}}, t_{\text{tail}}]$

$x \sim \mathcal{N}(0, I)$
for $k = 1, \ldots, K$ **do**
  $t \leftarrow t_k$; $\lambda_i \leftarrow \big[1 + (1{-}t)\,\widehat{\nabla \cdot v^i}(t, x)/d\big]^{-1}$ for $i \in \{A, B\}$ ▷ one Hutchinson probe; frozen within the slice
  for $s = 1, \ldots, S$ **do**
    $m^i \leftarrow x + (1{-}t)\,v^i(t, x)$, $s^i \leftarrow (t\,v^i(t, x) - x)/(1{-}t)$ ▷ Lemmas 1, 3; t capped at 0.995
    $\bar{\Lambda} \leftarrow \lambda_A + \lambda_B - t$, $b \leftarrow \lambda_A m^A + \lambda_B m^B - x$, $q \leftarrow \lambda_A \|m^A\|^2 + \lambda_B \|m^B\|^2 - \|x\|^2/t$
    if $t < t_{\text{cut}}$ and $\bar{\Lambda} > \rho$ **then**
      $\hat{g}_t \leftarrow \nabla_x \big[\frac{t}{2(1-t)^2}\big(\|b\|^2/\bar{\Lambda} - q\big)\big]$ with λ_A, λ_B held fixed ▷ frozen-covariance gradient of (5)
      $\hat{g}_t \leftarrow \hat{g}_t \cdot \min\{1,\ \|s^A + s^B\|/\|\hat{g}_t\|\}$ ▷ trust-region clip
    else
      $\hat{g}_t \leftarrow 0$ ▷ mask of Prop. 1, or past the cutoff
    end if
    $x \leftarrow x + \eta(t)\,(s^A + s^B + \hat{g}_t) + \sqrt{2\eta(t)}\,\xi$, $\xi \sim \mathcal{N}(0, I)$
  end for
end for
tail: integrate $\dot{x} = (m^{AB} - x)/(1{-}t)$ from t_{end} to t_{tail} (Euler), with $m^{AB} = b/\max(\bar{\Lambda}, \rho)$ and $\lambda_A, \lambda_B, b, \bar{\Lambda}$ recomputed at every step
return x

Diversity, FID, and step-size controls. MNIST diversity: mean pairwise L_2 over 512 samples is 13.3 (CompFM) vs. 15.2 (plain) vs. 16.3 (real eights); NN coverage 99.6% (CompFM) and 99.8% (plain). Folder-FID (n=512, against the real intersection class; corrected vs. plain): CIFAR 246 vs. 399, MNIST 254 vs. 276 (anneal-only protocol), FashionMNIST 307 vs. 295; at n=512 we read folder-FID as corroboration only where the gap is large. Step-size-scaled MNIST controls (plain path, 100 slices): at the baseline diffusion time, $10{\times}\eta/5$ and $40{\times}\eta/20$ stay at the 2-step baseline (78.6±2.6 / 78.6±1.1%); at the 10-step point's total time, $50{\times}\eta/5$ nearly matches the coarse point (92.2±0.5 vs. 94.0±1.1%); at the 40-step point's, $200{\times}\eta/5$ deepens the collapse (37.9±3.9 vs. 69.4±2.3%); coarse stepping partially masked the decline but cannot explain it.

Estimator ablation, guards, and cost. Table 5 crosses frozen vs. full gradients with scalar vs. full covariance on the oracle suite (corrected ULA @480, right-mass, 5 seeds). The gradient axis matters at oracle level: on M2 the frozen gradient forfeits most of the gain (0.953 scalar / 0.899 full-matrix vs. 0.994 full-gradient; plain 0.902), and on A3 it keeps a third; freezing hurts less under scalar covariance, whose weaker x-dependence gives the dropped $\nabla_x \Sigma$ terms less to carry. The scalar covariance's relative std across marginal-path points is 0.21–0.30 on MNIST vs. 0.67–0.87 on M2 at mid-t, a $\sim 3\times$ gap that correlates with, but does not prove, the benign behavior on images; numerical proximity to the full gradient there is not measured. All oracle results use full gradients,

the learned-flow gate uses finite differences of the full quadratic, and all image results use frozen-scalar. Mask-off: on M4 the oracle sampler crashes (Cholesky on the non-PD fusion); on M2 the certificate is globally positive and the mask vacuous; on MNIST the mask never binds ($100.0\pm0.0\%$ mask-off). The Σ-cache and the probe count make no measurable difference there (100 slices $\times$ 2 steps, 3 seeds: $99.9\pm0.1\%$ with the cache off, 100.0% with four probes), whereas the trust-region clip is what guards against the amplified conversion noise of Rem. 2. Cost: 4.7–$4.8\times$ wall-clock per anneal at the deployed setting ($5.0\times$ per slice, one probe, per-slice cache; the ratio varies with steps per slice).

Table 5: Estimator ablation: the corrected anneal at 480 calls under the four gradient and covariance combinations (5 seeds). The deployed image-side setting is frozen-∇, scalar.

right-mass	plain	full-∇, full-Σ	full-∇, scalar	frozen-∇, full-Σ	frozen-∇, scalar
M2 (true .9957)	.902±.003	.994±.001	.994±.001	.899±.003	.953±.003
A3 (true .9773)	.920±.008	.973±.003	.973±.003	.937±.006	.937±.006
M1 (true .4995)	.498±.007	.500±.007	.496±.007	.498±.007	.496±.006

Table 6: Operator family on the MNIST flows A (threes and eights, $80/20$) and B (eights and nines): audited target-class fraction (%, 5 seeds, n=512). Annealed methods use 100 slices $\times$ 2 steps, the velocity combination a 200-step ODE. Brackets: remaining eights for the negation. Seed-0 grids in Fig. 9.

operator (α, β)	target	corrected	summed scores	velocity comb.
product $(1, 1)$	eights	$\mathbf{100.0\pm0.1}$	79.5 ± 2.3	54.2 ± 2.4
negation $(1, -0.3)$	threes	$\mathbf{97.2\pm0.8}$ [0.2]	78.8 ± 1.2 [10.8]	84.8 ± 1.5 [11.7]
geometric $(0.5, 0.5)$	eights	$\mathbf{99.3\pm0.5}$	63.1 ± 3.0	53.2 ± 3.5

Table 7: Conditional conjunction on colored MNIST: joint accuracy (digit and color both correct, %, 5 seeds, n=512). Flows: 32-channel class-conditional UNets (10 digits; 3 colors), 8k steps; protocol as for MNIST (100 slices $\times$ 2 steps, no tail); audit: a digit CNN trained on the colored data and the dominant stroke channel. Seed-0 grids in Fig. 9.

sampler	red eights	green threes
corrected	$\mathbf{96.2\pm0.8}$	$\mathbf{98.5\pm0.7}$
summed scores	87.1 ± 1.5	77.8 ± 1.5
velocity averaging	38.5 ± 2.7	42.2 ± 1.2
digit-conditional flow alone	38.0 ± 1.2	33.6 ± 1.9
color-conditional flow alone	10.3 ± 1.4	12.2 ± 1.5

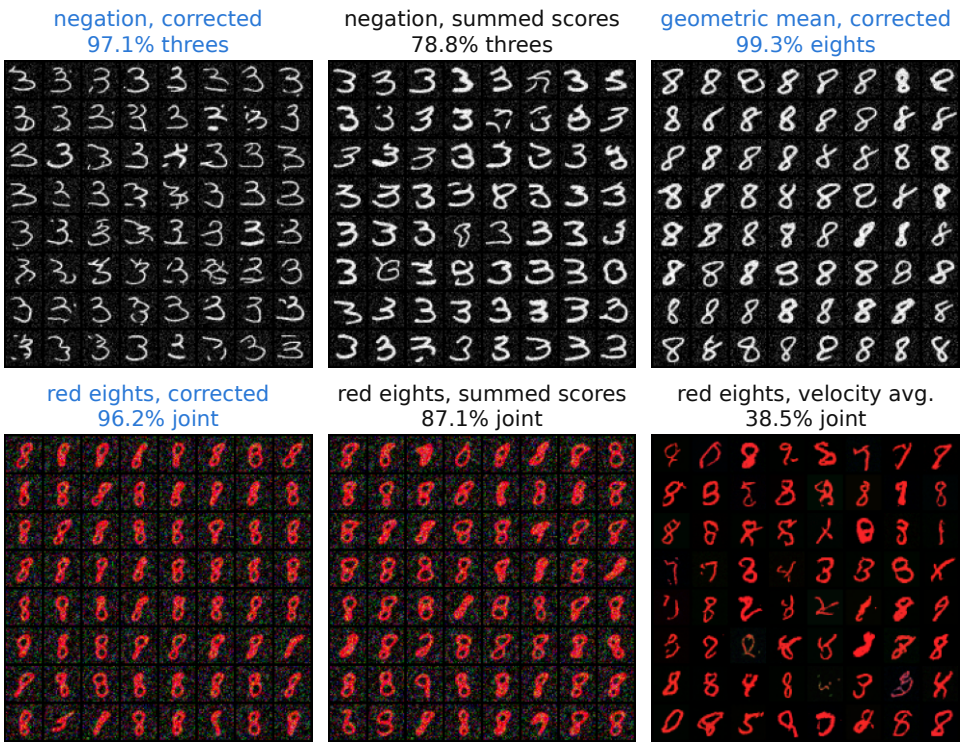


Figure 9: Seed-0 sample grids for Tables 6 and 7. Top: negation with exponent -0.3 on B, under the corrected path (predominantly threes) and under summed scores (eights remain), and the geometric mean under the corrected path (eights). Bottom: the conjunction of the digit-eight and red-color conditionals under the corrected path, summed scores, and velocity averaging.

C THE SMC LAYER AND THE FKC COMPARISON

For a deterministic drift ϕ_t transporting particles while targeting the unnormalized curve $\gamma_t = (p_t^A)^\alpha (p_t^B)^\beta \hat{R}_t$, the importance weight obeys $\frac{d}{dt}\log w = \partial_t \log \gamma_t + \phi_t \cdot \nabla \log \gamma_t + \nabla \cdot \phi_t$ (difference of the material derivatives of $\log \gamma_t$ and the Liouville density). Our implementation uses the discrete annealed-importance form (Neal, 2001): at slice boundaries $\log w \mathrel{+}= \log \gamma_{t_{k+1}}(x) - \log \gamma_{t_k}(x)$, with ULA as the forward kernel and systematic resampling below an effective sample size (ESS) of $0.5N$. With kernels that leave the intermediate targets invariant this scheme is consistent (Neal, 2001; Del Moral, 2004). Our deployed layer is not such a scheme, and we claim no consistency for it; its gaps differ by path. *Initialization* (both paths): the kernel factor keeps its explicit x-dependence at $t \to 0$: the $-\frac{\kappa}{2} c^\top \Lambda_k c$ term of (5) equals $-\kappa\|x\|^2/2(1-t)^2 \to \log q_0(x)^\kappa + \text{const}$, so $\hat{R}_0(x) \propto q_0(x)^{1-\alpha-\beta}$, exactly as for the exact R_0 (§3); numerically $\|\nabla \log \hat{R}_t - x\|_\infty \approx 4.6t$ on M2. Hence the *corrected* weight curve is correctly anchored, $\gamma_0^{\text{corr}} \propto q_0^2 \hat{R}_0 \propto q_0$, and starting from q_0 at t_0=0.02 omits only an $O(t_0)$ increment, provided the gate is open there. Whether it is depends on the benchmark: on M2 the certificate clears the gate's hard floor of 0.1 (unscaled precision units) at all 4096 sampled marginal-path points at t_0 (min. eigenvalue 0.90), and gated and ungated $\log \hat{R}_t$ agree to $< 10^{-6}$ on these points, supporting the anchoring numerically. On M4 the diffuse trimodal priors give eigenvalues $\approx$0.05 < 0.1 on all sampled points, so the sigmoid gate suppresses (gate $\approx$0.27) and the floored reconstruction distorts the correction there. At the finite t_0, the target is exactly $p_{t_0}^A p_{t_0}^B \cdot \hat{R}^{\text{gated}}$. The suppressed correction no longer guarantees anchoring to q_0, so the corrected weight curve can retain an $O(1)$ initial mismatch on certificate-poor benchmarks. This compounds the mid-t target mismatch below and is consistent with the measured M4 corrected+SMC failure (allocation TV 0.154 ± 0.034 vs. 0.035 weight-free). The *plain* curve, in contrast, has $t \to 0$ limit $\gamma_0^{\text{plain}} \propto q_0^2 \propto \mathcal{N}(0, \frac{1}{2} I)$ (at the finite t_0 it is exactly $p_{t_0}^A p_{t_0}^B$, inheriting the same $O(1)$ mismatch). This is an initial weight $\log \gamma_{t_0} - \log q_0$ that we omit, fol-

lowing common practice; it is not repaired by $N \to \infty$ or by shrinking the step size. Empirically the plain rows of Table 8 nevertheless reach 4×10^{-4} at large budgets, an observation consistent with (but not proof of) fast early mixing at near-Gaussian targets. *Target mismatch* (corrected path only): the weight increments use a smoothly gated $\hat{R}_t$ while the dynamics use the hard mask and clip, so the kernels do not leave the weight targets invariant at mid-t even in continuous time. *Discretization* (both paths): unadjusted fixed-step ULA carries its stationary bias. A consistent version would initialize each path with $\log \gamma_{t_0}(x) - \log q_0(x)$, match dynamics to a differentiable target (or Metropolis-adjust), and use exact transition ratios; we leave this to future work. Both paths share the kernel budget and the ULA discretization caveat, and the remaining gaps are path-specific as itemized above, so Table 8 is a matched-budget comparison of two approximate pipelines, not of two consistent estimators. All terms are computable from black-box flows: scores via Lemma 3, $\partial_t \log p_t^i = -\nabla \cdot v^i - v^i \cdot s^i$ by the continuity equation with Hutchinson divergence (Hutchinson, 1989), and $\hat{R}_t$ in closed form. The 2D protocol fixes a total particle budget (4096) split into interacting populations of size N with forced final resampling and pooled evaluation (5 seeds). We additionally cross-validated our implementation against the official FKC release. We ported the product sampler of Skreta et al. (2025a) verbatim (their VP schedule, product-SDE drift, continuous log-weight rate $d \log w = dt\,[2\, d \log \alpha_t / dt + (2\zeta_t/\sigma_t)\langle s^1, s^2 \rangle]$, where ζ_t is their schedule coefficient, written γ_t in their paper and unrelated to our curve γ_t, and per-step systematic resampling) and ran it with oracle scores. It reproduces their published toy behavior and, on M2 with a global population, attains allocation error 0.0005 ± 0.0004 (5 seeds, 1000 steps), matching the plain+SMC rows of Table 8. Resample-move dynamics yield unique fractions of 0.82–0.94 across the SMC rows. At 48 NFE, terminal SNIS yields unique fractions of 11.7% on the plain path and 18.2% on the corrected path.

Results. With exact densities (2D oracle), SMC reweighting on the plain path fixes allocation estimates at large budgets (Table 8). First, without any weights the corrected path reaches allocation error 0.045 ± 0.004 at 48 NFE (5 seeds; the unweighted plain path gives 0.319 ± 0.003 at 48 and 0.094 ± 0.003 at 480); a single terminal self-normalized importance (SNIS) reweight brings it to 0.0016 ± 0.0017, on par with $N{=}256$ interacting SMC (0.0019). SMC is unavailable when samples must be generated independently. Second, corrected-path SMC outperforms plain-path SMC at every N at tight budgets (0.0009 vs. 0.0019 at $N{=}256$), and its importance weights survive $8\times$ better without resampling (Fig. 2c). Third, on MNIST with weights estimated from network divergences, FKC needs a global 512-particle population and 100 slices to match our weight-free sampler; at population 64 it drops to $87.8 \pm 12.6\%$, and at 25 slices it is unreliable ($66.7 \pm 57.7\%$) where ours reaches $97.5 \pm 0.4\%$ (App. C). The closed-form correction therefore makes composition work per sample, at small populations and at tight budgets, and the two methods compose (Table 8, App. C).

Table 8: M2 oracle allocation error of pooled populations (4096 particles in total; u = unique fraction, shown where informative), using oracle densities to evaluate the weight increments. The MNIST comparison under estimated weights is in App. C.

budget	path	SNIS only	SMC $N{=}16$	SMC $N{=}64$	SMC $N{=}256$
48 NFE	plain	.0037±.0071 (u .12)	.0098±.0055	.0044±.0048	.0019±.0010
	corrected	**.0016±.0017** (u .18)	.0037±.0034	.0029±.0022	**.0009±.0007**
480 NFE	plain	.0010±.0006	.0004±.0004	.0003±.0002	.0005±.0003
	corrected	.0009±.0003	.0014±.0009	.0017±.0008	.0010±.0007

D A LIKELIHOOD AUDIT OF THE CIFAR-SCALE INTERSECTION

The CIFAR gap (§6.3) admits two readings: the sampler fails to track its target, or the target itself, the product of the *learned* densities, places its mass away from the target class. The two are separated by model log-likelihood estimates: for a flow ODE $\dot{x} = v(t, x)$, $\log p_1(x) = \log p_{t_0}(x_{t_0}) - \int_{t_0}^{1} \nabla \cdot v \, dt$ along the backward trajectory (Grathwohl et al., 2019), which we integrate with a fixed-grid 100-step RK4 and Hutchinson divergence estimates (8 Rademacher probes drawn once per trajectory and held fixed across integrator stages). The implementation is validated against analytic GMM oracles (median integration error $< 10^{-4}$ nats; estimator z-scores calibrated against exact divergences); on CIFAR the median probe noise is 31 nats and doubling the integrator steps moves medians by 15 nats, both well below the approximately 1000-nat between-class median differences reported below. Real

images are not dequantized, so absolute values are not comparable to bits-per-dim; every conclusion below is a same-protocol between-group comparison, with group medians and bootstrap standard errors.

No detectable sampler-side bias. For corrected-sampler outputs under the deployed protocol (seed 0, n=512), the ship-audited group (n=327) and the truck-audited group (n=65) are statistically indistinguishable under the learned product $\log p^A + \log p^B$ (median gap 33 nats against a 3 s.e. threshold of 102). The test does not attribute the selection of ships over trucks to the sampler; under the learned product the target itself places its mass there.

An inverted veto. Define the veto gap $\Delta^A := \operatorname{med} \log p^A(\text{truck}) - \operatorname{med} \log p^A(\text{ship})$ over real held-out images (512 per class; Δ^B analogously against automobiles). Flow A, trained only on automobiles and trucks, assigns real ships a median log-likelihood 986 ± 103 nats higher than real trucks. Flow B, trained on trucks and ships, prefers ships by $\approx$1036 nats (as expected, since they are in its support), so product mass can reach trucks only if A's veto exceeds roughly $+1000$ nats: a $\approx$2000-nat swing from the measured value.

Robust to capacity and duration. Retraining both flows at $4\times$ capacity (128-channel UNets) leaves the gaps unchanged at matched steps and at doubled duration (Table 9). The stronger flows are nevertheless visibly better single-model generators: flow A's own samples audit at $78/20$ auto/truck under the paper's auditor, matching its $80/20$ training mixture, at confidence 0.97. Composition purity does not improve with the stronger flows: truck 7.0% under the paper's auditor at seed 0 (2.0% under an independent 95.1%-accuracy auditor of the same architecture; plain 0%; deployed weak-flow reference $11.0\pm2.1\%$). Purity tracks the veto gap, not single-model sample quality.

Table 9: Veto gaps (nats; medians over matched seed-0 subsets of 256 real test images per class, bootstrap s.e.) across model capacity and training duration (training halted once the curve stayed flat through the 100k mark). The class-selectivity of the learned densities does not move.

flows	Δ^A (truck−ship)	Δ^B (truck−auto)
64ch, 50k steps (deployed)	-812.5 ± 150.2	$+213.0 \pm 144.9$
128ch, 50k steps	-809.9 ± 142.2	$+216.1 \pm 137.7$
128ch, 100k steps	-854.9 ± 152.3	$+233.1 \pm 151.9$

Interpretation. This is the within-dataset analogue of the complexity bias of pixel-space likelihoods (Nalisnick et al., 2019; Kirichenko et al., 2020): log-likelihood is dominated by local smoothness statistics (ships contain large sky and water regions), robustly to capacity and duration. The boundary therefore concerns pixel-space density products as composition targets and is shared by any sampler that reaches them faithfully; it is stratified by data complexity, and MNIST and FashionMNIST sit on the favorable side. The likelihood instrument itself is covered by closed-form golden tests.

E OT-COUPLED FLOWS BREAK THE KERNEL STRUCTURE

Lemma 2 requires $x_0 \perp x_1$. For minibatch-OT couplings (Tong et al., 2024; Pooladian et al., 2023) the conditional $\pi(x_0|x_1)$ is not q_0, the shared kernel disappears, and Theorem 1 acquires coupling-ratio terms that are not available from the trained field. Empirically (M2, learned MLPs trained identically except for the coupling): with I-CFM couplings the corrected anneal reaches W_2=0.75 at 480 NFE, while with OT-CFM couplings *both* the plain and corrected samplers degrade to $W_2\approx1.83$–1.88 at all budgets. The score–velocity conversion itself (Lemma 3) is coupling-dependent, so methods that route through it, ours included, lose their justification under these couplings.