When is Compositional Generation Feasible?

Distributional Estimation Error and Inference-time Approximation Error in Diffusion Models

Anonymous authors
Paper under double-blind review

Abstract

The task of *compositional generation* involves using a conditional generative model, trained only on a subset of the possible conditions, to produce samples from compositionally-defined target distributions such as a geometric combination of the source distributions. In this work, we explore what makes this challenging by separating out several sources of error and analyzing their interplay. To begin, we show that even when source distributions are estimated with arbitrarily small error, their composition may be arbitrarily far from the true composition. This leads us to derive sufficient conditions to prevent error blowup, and we propose a metric based on the resulting error-sensitivity bounds which can separate error-sensitive compositions from stable ones without access to ground truth distributions. Then, we carefully design experiments with diffusion models for both synthetic and realistic data to analyze the interaction between distributional estimation error and inference-time approximation error. In particular, while recent methods can reduce *inference-time approximation error* in principle, our results show that *distributional estimation error* has a more catastrophic effect on performance when the target distribution is out-of-distribution with respect to the sources.

1 Introduction

The space of distributions that we would like to sample is often far larger than the set of "source" distributions seen during training. We want models that can imagine arbitrary combinations of concepts (e.g. "a *living room* with a *white couch*, a *black chair*, two *paintings*, a *floor lamp* and nothing else"), even if the training data includes only some combinations of these concepts. This is particularly true when data are derived from experiments, e.g., in biology, the space of possible experiments is combinatorially large (covering all possible combinations of molecules, gene knockouts, cell types, assays, etc.), but experimental budgets are finite. Hence, there is significant interest in building models that can predict the outcomes of unseen experiments (Lotfollahi et al., 2023; Roohani et al., 2024; Wang et al., 2024; Adduri et al., 2025; Wenkel et al., 2026; Bunne et al., 2024; Noutahi et al., 2025).

To extrapolate to new combinations, methods for *compositional generation* rely on strong inductive biases, e.g., assuming the existence of some latent space in which the effects are additive. For example, if two perturbations a and a' are *causally separable*, one can show the double-perturbation distribution $P(x \mid \text{do}(a), \text{do}(a'))$ can be expressed as a geometric combination of the control distribution $P(x)$ and the single-perturbation distributions $P(x \mid \text{do}(a))$ and $P(x \mid \text{do}(a'))$ (Wang et al., 2023; Xu et al., 2024)[1]. In principle, such results imply significant reductions in the amount of data necessary to succeed at these difficult generation tasks. In practice, however, even if the compositional distribution $P(x \mid \text{do}(a), \text{do}(a'))$ can be expressed in terms of source distributions, this does not imply that we can efficiently sample from $P(x \mid \text{do}(a), \text{do}(a'))$: distribution-level identities do not necessarily translate into computationally and statistically efficient algorithms. In this paper, we focus on procedures for composing *conditional diffusion models*, given their practical relevance and powerful generative capabilities (Dhariwal & Nichol, 2021).

[1] In practice, causal separability is not expected to hold for *all* combinations of biological perturbations However, if we can accurately characterize $P(x \mid \text{do}(a), \text{do}(a'))$ when separability *does* hold, then biologists could focus experimental budgets on combinations of perturbations that are not expected to be separable.

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.

Notably, geometric combinations at the distribution level translate into linear combinations at the score level. Hence, a popular heuristic for compositional generation is to linearly combine the scores (Liu et al., 2022) during denoising. However, it is increasingly recognized that this naïve approach introduces a fundamental source of error: in general, adding noise to the distributions breaks the linear relationship between their score functions. To account for this issue, recent correction algorithms (Skreta et al., 2025; Xie et al., 2026; Ren et al., 2026) explicitly correct for this error. Despite these advances, we highlight a separate issue which also poses significant problems for composition. We find that in many settings of interest (those in which the target distribution places large mass in low-density regions of the source distributions), the distributional estimation error, i.e., the error propagated from estimation of the source distributions, becomes the dominant factor. Moreover, in these cases, we find that correction reinforces these errors, further degenerating sample quality.

Contributions Building on prior work, we first introduce a *concrete formulation of the compositional generation task*, show that such compositions are *closely connected to causal representation learning*, and *formalize two sources of error* that make this task difficult (Sec. 2 and App. B). Then, we *derive a bound which decouples estimation error from compositional sensitivity to that error*, that admits a practical estimator, and *characterize the interplay between these terms and inference-time approximation error* (Sec. 3). Based on these insights, we perform careful experimentation to disentangle distributional estimation error from inference-time approximation error in synthetic and practical settings (Sec. 4). These experiments reveal that both sources of error lead to compositional failure within distinct, partially overlapping regimes, and that within the overlap, distributional estimation error often dominates attempts to correct for inference-time approximation error.

Notation Given a measure space $\mathcal{X}$, we let $\mathcal{P}(\mathcal{X})$ denote the set of probability distributions on $\mathcal{X}$. Throughout, we assume that all distributions have densities and that these densities are positive, ensuring that we avoid division by zero.[2] With a typical abuse of notation, given a distribution $P \in \mathcal{P}(\mathcal{X})$, we also write P for its associated density. Finally, we let $\mathbf{e}_k$ denote the kth basis vector.

2 Setup: Geometrically weighted compositions

Mathematically, we consider a ***perturbation space*** $\mathcal{A}$ and an ***outcome space*** $\mathcal{X}$. Each perturbation $a \in \mathcal{A}$ is associated with some ground truth distribution $P^a \in \mathcal{P}(\mathcal{X})$ over outcomes. In practice, we may only observe outcome data for a small subset $\mathcal{A}_o \subseteq \mathcal{A}$ of the overall perturbation space, and we may wish to *extrapolate* to an unseen perturbation $a^* \in \mathcal{A} \setminus \mathcal{A}_o$. Under a variety of well-motivated theoretical assumptions, the target distribution P^{a^*} can often be *identified* from the observed distributions $\mathbf{P} = (P^a)_{a\in\mathcal{A}_o}$, i.e., $P^{a^*} = f(\mathbf{P})$ for some known *identification formula* f.

Example 1. *Suppose a scientist can apply K drugs at varying dosages and observe p outcomes ($\mathcal{A} = \mathbb{R}^K_{\geq 0}$, $\mathcal{X} = \mathbb{R}^p$). Since experiments are costly, they may only have data for the control condition and for single drugs at fixed dosages, i.e., $\mathcal{A}_o = \{\mathbf{0}\} \cup \{\mathbf{e}_k\}_{k\in[K]}$, and wish to extrapolate to a double-perturbation, e.g. $a^* = \mathbf{e}_1 + \mathbf{e}_2$.*

One may expect that perturbations have targeted effects: in causality, such effects can be modeled as mechanism changes *or* soft interventions *(Squires & Uhler, 2023). For example, assume $\mathcal{X}$ can be split into upstream and downstream components $\mathcal{X}_1$ and $\mathcal{X}_2$, so that $P^0(x) = P^0(x_1)P^0(x_2|x_1)$ is the observational distribution, while $P^{\mathbf{e}_1}(x) = P^{\mathbf{e}_1}(x_1)P^0(x_2|x_1)$ and $P^{\mathbf{e}_2}(x) = P^0(x_1)P^{\mathbf{e}_2}(x_2|x_1)$ are interventional distributions. Under the assumption of* independent causal mechanisms *(Schölkopf et al., 2021), interventions act modularly, implying that $P^{a^*}(x) = P^{\mathbf{e}_1}(x_1)P^{\mathbf{e}_2}(x_2|x_1)$. Hence P^{a^*} can be written as a function of P^0, $P^{\mathbf{e}_1}$, and $P^{\mathbf{e}_2}$, in particular, $P^{a^*}(x) = P^{\mathbf{e}_1}(x) \cdot P^{\mathbf{e}_2}(x) \cdot P^0(x)^{-1}$.*

In this work, we consider a more general form for f that is common in several settings, where the density of distribution P^{a^*} is expressed as a geometric mixture of the source distribution densities.

Definition 1. *Fix $\mathcal{A}_o \subseteq \mathcal{A}$. Given a collection of distributions $\mathbf{P} = (P^a)_{a\in\mathcal{A}_o}$ and a function $w\colon \mathcal{A}_o \to \mathbb{R}$, we say that w is a* valid weighting *if $f(x) = \prod_{a\in\mathcal{A}_o} P^a(x)^{w_a}$ is integrable, where w_a is shorthand for $w(a)$. For a valid weighting w, we define the* **weighted composition** *of $\mathbf{P}$ as*

$$\textsc{Comp}_w(\mathbf{P})(x) := \frac{1}{Z^w} \prod_{a\in\mathcal{A}_o} P^a(x)^{w_a}, \quad \text{where } Z^w := \int_{\mathcal{X}} \left(\prod_{a\in\mathcal{A}_o} P^a(x)^{w_a} \right) dx. \tag{1}$$

[2] In App. A, we formalize this assumption and discuss the formal interpretation of our equality expressions.

As shorthand, we write $\mathbf{P}^w := \text{Comp}_w(\mathbf{P})$. This geometric combination at the distribution level becomes a linear combination at the score level. Assuming $\log P^a$ is differentiable, the *Stein score functions* are $S^a(x) := \nabla_{\mathsf{x}} \log P^a(x)$ for $a \in \mathcal{A}_\text{o}$, and $S^w(x) := \nabla_{\mathsf{x}} \log P^w(x)$. Then, for any valid weighting w, we have $S^w(x) = \sum_{a \in \mathcal{A}_\text{o}} w_a S^a(x)$.

As an important special case, in line with Example 1, we often consider ***base-composed distributions***, where $0 \in \mathcal{A}_\text{o}$ is the index of the *base distribution*, and $w_a = 1 - \mathbb{1}_{a=0} \cdot |\mathcal{A}_\text{o}|$. This weighting gives

$$\mathbf{P}^w(x) = \frac{1}{Z^w} P^0(x) \prod\nolimits_{a \in \mathcal{A}_\text{o}} \frac{P^a(x)}{P^0(x)}, \tag{2}$$

the same form of composition studied by Bradley et al. (2025). As we describe in App. B, such compositions arise naturally in the field of *causal representation learning*; this framework extends the motivation from Example 1 to multiple perturbations, and to settings where $\mathcal{X}$ is not directly decomposable into causal components. We emphasize that this extends the notion of Bayes composition (Bradley et al., 2025; Liu et al., 2022; Du et al., 2023), where P^0 is a marginal distribution and the $P^a \in \mathbf{P}$ are conditional distributions, all arising from the same joint distribution.

The challenges of compositional generation With this motivation, we define the concrete task:

Task (Compositional generation for weighted compositions)**.** *Let d be some metric on the space of distributions $\mathcal{P}(\mathcal{X})$, and let w be a valid weighting for $\mathbf{P} = (P^a_\star)_{a \in \mathcal{A}_o}$. Suppose we are given a pre-trained conditional generative model $\widehat{\mathbf{P}} = (\widehat{P}^a)_{a \in \mathcal{A}_o}$ such that $d(\widehat{P}^a, P^a_\star) < \varepsilon_1$ for all $a \in \mathcal{A}_o$. As output, we wish to produce samples from some $\widetilde{P}$ such that $d(\widetilde{P}, \mathbf{P}^w_\star) < \varepsilon_2$.*

In particular, given accurate estimates $(\widehat{P}^a)_{a \in \mathcal{A}_\text{o}}$ of source distributions $(P^a_\star)_{a \in \mathcal{A}_\text{o}}$, we wish to approximately sample from a target distribution $P^w_\star$, defined via weighted composition. This task encounters two challenges. First, one encounters *inference-time approximation error*: for many model classes (e.g. diffusion models), one cannot efficiently sample from $\widehat{\mathbf{P}}^w$, only some proxy $\widetilde{P} \approx \widehat{\mathbf{P}}^w$. Second, we encounter *propagated distributional estimation error*: since we are composing estimated distributions $\widehat{\mathbf{P}} \neq \mathbf{P}_\star$, those estimation errors propagate to $\widehat{\mathbf{P}}^w := \text{Comp}_w(\widehat{\mathbf{P}}) \neq P^w_\star$. In particular, given any metric d on the set of probability distributions $\mathcal{P}$, the triangle inequality gives

$$d(\widetilde{P}, \mathbf{P}^w_\star) \leq \underbrace{d(\widetilde{P}, \widehat{\mathbf{P}}^w)}_{\text{inference-time approximation error}} + \underbrace{d(\widehat{\mathbf{P}}^w, \mathbf{P}^w_\star)}_{\text{distributional estimation error}}. \tag{3}$$

As we discuss in App. C, recent methods such as *Feynman-Kac correction* (Skreta et al., 2025) provide elegant solutions to reduce the inference-time approximation error $d(\widetilde{P}, \widehat{\mathbf{P}}^w)$ when composing diffusion models. Although reducing this error is important, we find that the distributional estimation error $d(\widehat{\mathbf{P}}^w, \mathbf{P}^w_\star)$, often overlooked in previous works, can be much more important in certain settings: small errors in the estimated source distributions $\widehat{P}^a$ can blow up into large errors in the implied target distribution $\widehat{\mathbf{P}}^w$, due to ill-conditioning of the composition operator Comp_w. Specifically, we consider settings where typical samples from $\mathbf{P}^w$ lie in low-density regions of the source distributions; for convenience, in these cases, we say that $\mathbf{P}^w$ is *out-of-distribution (OOD)* with respect to the source distributions.[3]

3 Theoretical Insights

In this section, we analyze how the highlighted sources of error in compositional generation can lead to catastrophic failure. For clarity, we focus on base compositions with $\mathcal{A}_\text{o} = \{0, a_1, a_2\}$.

3.1 Distributional Estimation Error

As a preliminary, we define the *relative error* for each source as $u_i(x) := (\widehat{P}^i(x) - P^i(x))P^i(x)^{-1}$ for $i \in \{0, a_1, a_2\}$, where we have that $\|u_i\|^2_{L^2(P^i)} = \chi^2(\widehat{P}^i \| P^i)$ (S. Agapiou & Stuart, 2017). This

[3]To justify this terminology, consider the example of diffusion models: the estimated score functions are trained on $\mathbf{P} = (P^a)_{a \in \mathcal{A}_o}$ but are evaluated on $\mathbf{P}^w$, similar to the setting considered in the field of *OOD generalization* (David et al., 2010; Kpotufe & Martinet, 2018; Canatar et al., 2021).

is not the only kind of distributional error we consider but, as we will see, it permits useful bounds. First, we show that arbitrarily small source-wise error can lead to arbitrarily poor compositions:

Proposition 1 (Arbitrarily unstable composition). *There exist sequences of one-dimensional Gaussian source distributions* $(P_n^0, P_n^{a_1}, P_n^{a_2})$ *and* $(\widehat{P}_n^0, \widehat{P}_n^{a_1}, \widehat{P}_n^{a_2})$ *whose source-wise errors vanish simultaneously in TV, KL, and* W_2*, and where* $\|u_{i,n}\|^2_{L^2(P_n^i)} \to 0$ *for all* $i \in \{0, a_1, a_2\}$*, while the corresponding base compositions* $\widehat{\mathbf{P}}_n^w \propto \widehat{P}_n^{a_1} \widehat{P}_n^{a_2} (\widehat{P}_n^0)^{-1}$ *satisfy*

$$\mathrm{TV}(\mathbf{P}_n^w, \widehat{\mathbf{P}}_n^w) \to 1, \qquad D_{\mathrm{KL}}(\mathbf{P}_n^w \| \widehat{\mathbf{P}}_n^w) \to \infty, \qquad W_2(\mathbf{P}_n^w, \widehat{\mathbf{P}}_n^w) \to \infty.$$

The proof and a similar proof for Bayes composition are given in Apps. G.2 and G.12, respectively. This result motivates conditions which prevent such pathological error blowup. Accordingly, we derive a local bound under relatively mild assumptions that separates source error from the sensitivity of the composition to that error.

Theorem 1 (First-order sensitivity bound). *For* $s \in [0, 1]$*, consider the convex interpolation* $P_s^i = P^i(1 + s u_i) = (1 - s)P^i + s\widehat{P}^i$ *and let* P_s^w *denote the corresponding base composition whenever it is valid. Define source-to-composition ratios* $R_i(x) = \mathbf{P}^w(x) P^i(x)^{-1}$ *and*

$$\kappa_i := \left(\mathbb{E}_{P^i}[R_i(x)^2]\right)^{1/2} = \sqrt{1 + \chi^2(\mathbf{P}^w \| P^i)}.$$

Assuming P_s^w *is well-defined for all sufficiently small* $s \geq 0$ *and* $s \mapsto P_s^w$ *is right-differentiable in* L^1 *at* $s = 0$*, then as* $s \downarrow 0$*,*

$$\mathrm{TV}(\mathbf{P}^w, P_s^w) \leq s \sum\nolimits_{i \in \{0, a_1, a_2\}} \kappa_i \|u_i\|_{L^2(P^i)} + o(s).$$

This result is tight up to constant factors, can be trivially extended to general weighted compositions, and can be strengthened to a global bound, see Apps. G.3 and G.9 for details. Importantly, Theorem 1 separates the learned distributional error from the *sensitivity* to that error in a composition. Thus, when the κ_i are small for the composition of interest, then learning the source distributions to small relative error is (locally) sufficient for effective composition.

3.2 Inference-time approximation error

Now, we turn to the first term in Equation (3), focusing on diffusion models, where exact sampling from $\widehat{\mathbf{P}}^w$ is non-trivial. As we will see, naïve composition methods exhibit a tradeoff. On one hand, they can introduce arbitrarily large inference-time approximation error. On the other hand, in certain well-structured settings, naïve methods are less sensitive to distributional estimation error; the tradeoff between these two sources of error is reflected in Sec. 4.

We start with a minimal review to establish the necessary background and notation. Diffusion models operate on noised versions of the data distribution: letting Noise_t denote the noising operator represented by the noising SDE $\mathrm{d}x_t = u_t(x_t)\mathrm{d}t + \sigma_t \mathrm{d}W_t$, we write the noised distributions as $P_t^a := \mathrm{Noise}_t(P^a)$, and we write the corresponding score functions as $S_t^a(x) := \nabla_{\mathsf{x}} \log P_t^a(x)$.[4] Then, samples from P_t^a can be generated by integrating the reverse-time SDE $\mathrm{d}x_t = \left(u_t(x_t) - \sigma_t^2 S_t^a(x_t)\right)\mathrm{d}t + \sigma_t \mathrm{d}\overline{W}_t$ for $t = 1$ to $t = 0$ (Song et al., 2021).

To sample from the target $\mathbf{P}^w$, the denoising process instead requires $S_t^w(x) := \nabla_{\mathsf{x}} \log \mathbf{P}_t^w(x)$ for $\mathbf{P}_t^w := \mathrm{Noise}_t(\mathbf{P}^w)$. It is common to approximate $S_t^w(x)$ with the ***naïvely composed score*** as an additive heuristic, $S_t^{w,\mathrm{nv}} := \sum_{a \in \mathcal{A}_\circ} w_a S_t^a(x)$ (see e.g. classifier-free guidance (Ho & Salimans, 2022) and compositional visual generation (Liu et al., 2022)). However, while weighted composition ensures that $S_t^{w,\mathrm{nv}} = S_t^w$ at $t = 0$, noising and composition do not generally commute, so in general $S_t^w \neq S_t^{w,\mathrm{nv}}$ for intermediate t. As noted in prior works (Bradley & Nakkiran, 2024; Skreta et al., 2025), this discrepancy produces inference-time approximation error; in App. G.4, we provide an analytically tractable example and prove that this error can *also* be arbitrarily large.

In contrast to naïve methods, Feynman-Kac Correctors (Skreta et al., 2025) derive importance weights that exactly correct this discrepancy along the denoising trajectory. To realize these corrections in

[4] Here, we use the standard diffusion model convention that $t = 0$ corresponds to data and $t = 1$ corresponds to noise; W_t denotes the Weiner process and $\overline{W}_t$ denote the reverse Weiner process.

practice, Skreta et al. (2025) propose a Sequential Monte Carlo scheme that maintains K weighted particles updated according to a residual term $g_t(x)$, and resamples according to these accumulated weights. We refer the reader to App. C for further background on diffusion and FKC.

FKC mitigates the bias from inference-time approximation error, but can be more susceptible to variance from distributional estimation error. To see this, we consider Factorized Conditionals, a special case of base-composed distributions that can be interpreted as disconnected components in a causal graph. Adapting from Bradley et al. (2025), we say a collection of distributions $\mathbf{P} = (P^0, P^{a_1}, \dots P^{a_k})$ over $\mathbb{R}^n$ are ***Factorized Conditionals*** if there is a partition $M_0, M_1, \dots M_k$ of $[n]$ such that

$$P^{a_i}(x) = P^{a_i}(x_{M_i})P^0(x_{M_i^c}) \qquad \text{and} \qquad P^0(x) = P^0(x_{M_0}) \prod\nolimits_{i\in[k]} P^0(x_{M_i}), \tag{4}$$

where M_i^c denotes the set complement of M_i. Visually, this could correspond to placing objects which occur in disjoint regions of an image (e.g., couch and wall-mounted painting) against a shared background scene P^0 (e.g., empty room). If $\mathbf{P}$ are Factorized Conditionals, then *assuming zero score estimation error*, running the naïve denoising SDE correctly samples from $\mathbf{P}^w$. Bradley et al. (2025) show that this result can be generalized to feature space, where conditions could, for instance, orthogonally represent the style of an image versus its content, see App. E.1 for further details.

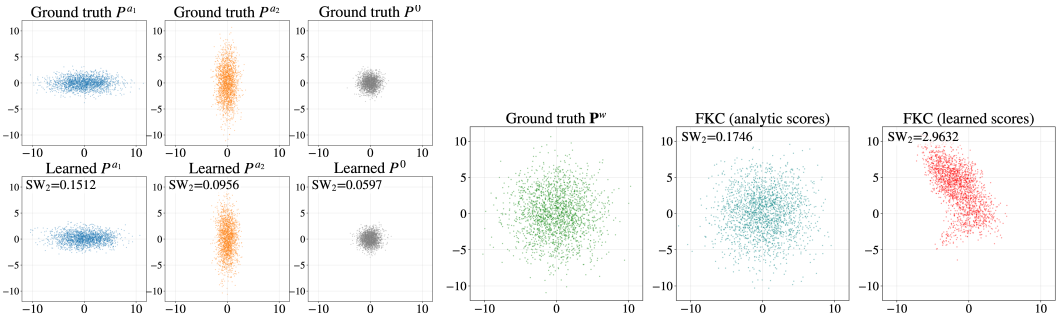
In fact, as we prove in App. G.5, if $\mathbf{P}$ are Factorized Conditionals, the Feynman-Kac Corrector log-weight drift $g_t(x)$ vanishes, and FKC coincides exactly with the naïve denoising SDE, since there is no inference-time approximation error to correct. However, with learned scores $\widehat{S}^a = S^a + \varepsilon^a$ due to estimation error, the cancellation that causes g_t to vanish no longer holds exactly, and g_t is driven entirely by estimation noise. Consequently, we might expect FKC to drift further from the true distribution rather than closer, and for this effect to worsen for out-of-distribution compositions, where ε^a is likely to be larger. This mechanism helps to explain our empirical observations in Sec. 4.

Furthermore, if $\mathbf{P}$ are (approximately) Factorized Conditionals, then exact composition can actually perform *worse* than naïve composition if the FC structure is not enforced during estimation. As a simple example, consider $P^0 = \mathcal{N}(0, I_2)$, $P^{a_1} = \mathcal{N}(0, \mathrm{diag}(n, 1))$, and $P^{a_2} = \mathcal{N}(0, \mathrm{diag}(1, n))$, so that $\mathbf{P}^w = \mathcal{N}(0, nI_2)$. Perturbing only $\widehat{P}^0$ to $\mathcal{N}(0, \mathrm{diag}(1+\varepsilon_1, 1+\varepsilon_2))$, we have $\widehat{\mathbf{P}}^w = \mathcal{N}(0, \mathrm{diag}(\tau_1, \tau_2))$ for $\tau_j = n(1+\varepsilon_j)/(1+(n+1)\varepsilon_j) = \Theta(1)$, whereas the naïve denoising ODE yields $\widetilde{P}^w = \mathcal{N}(0, \mathrm{diag}(\nu_1, \nu_2))$ for $\nu_j = n/(1+\varepsilon_j) = \Theta(n)$. Proofs of these facts and an approximate FC example are in App. G.4. Hence, naïve composition can perform a kind of regularization: the overall operation is less sensitive to small perturbations of the source distributions.

4 Experiments

The goal of our empirical evaluation is to disentangle the two highlighted sources of error in compositional generation: inference-time approximation error and distributional estimation error. We divide our tasks into In-Distribution (ID) settings, where the composed target $\mathbf{P}^w$ is well-represented within the training data sampled from the source distributions, and Out-of-Distribution (OOD) settings, where the composed target does not lie within the effective training support of the source distributions (a quantification of this property is provided in App. D). This allows us to control the estimation error. To separate out the effects of approximation error, we further adjust distributions to either be Factorized Conditionals or Non-Factorized Conditionals. In realistic, high dimensional settings such as image generation, it is not in general possible to determine factorization without access to ground-truth distributions. In App. E.2, we describe a practical heuristic that can be used instead.

In each experiment, we define two perturbation distributions P^{a_1} and P^{a_2} alongside a base distribution P^0. A conditional diffusion model $\widehat{\mathbf{P}}$ is trained to sample from these distributions, after which FKC is applied to this model to sample from the base-composed distribution (2). We control the strength of the Feynman-Kac correction by altering the number of simulated particles (K), and we exert further control over the estimation error by varying the number of training samples (N) or by using the analytic score functions. On synthetic experiments, we evaluate using the sliced Wasserstein-2 distance (SW_2) and Maximum Mean Discrepancy (MMD^2) compared to the ground truth base-composition. In the main text, we report mean SW_2, while App. K provides both metrics and reports standard deviation. Ablations with latent diffusion models and separate conditional models are presented in App. H, where similar trends hold. All experimental details are provided in App. I, and code is provided at anonymous.4open.science/r/compose-anonymous-A3E1/.



(a) Individual distributions (b) Composition

Figure 1: **The composition is accurately estimated with analytical scores, but fails with learned distributions due to out-of-distribution estimation error, even with Feynman-Kac correction.**

Table 1: **The relative magnitudes of inference-time approximation error and distributional estimation error are highly dependent on problem setting.** FKC is applied with K particles to a conditional diffusion model trained on N samples, or to the analytic score functions. Each value is the SW_2 between 5000 samples from $\mathbf{P}^w$ and 5000 empirical samples, averaged over 30 training runs.

(a) **Factorized conditionals + ID**

$K\backslash N$	100	1000	10000	analytic
1	0.3537	0.1089	0.0747	0.0354
4	0.4276	0.0974	0.0750	0.0376
16	0.4894	0.0956	0.0772	0.0351
64	0.5136	0.0937	0.0758	0.0356
256	0.5261	0.1010	0.0777	0.0346

(b) **Non-factorized conditionals + ID**

$K\backslash N$	100	1000	10000	analytic
1	0.4454	0.1234	0.0957	0.0655
4	0.5331	0.0969	0.0751	0.0412
16	0.5914	0.0901	0.0750	0.0353
64	0.6111	0.0905	0.0740	0.0357
256	0.6216	0.0963	0.0793	0.0333

(c) **Factorized conditionals + OOD**

$K\backslash N$	100	1000	10000	analytic
1	0.9658	0.5577	0.3948	0.1122
4	2.0688	1.6885	1.1773	0.1130
16	2.9177	2.7214	1.9699	0.1133
64	3.3210	3.3348	2.4983	0.1124
256	3.5097	3.6426	2.8210	0.1140

(d) **Non-factorized conditionals + OOD**

$K\backslash N$	100	1000	10000	analytic
1	1.0957	0.9154	0.8518	0.6375
4	2.1611	1.8306	1.3763	0.1764
16	3.0011	2.8830	2.0117	0.0981
64	3.4224	3.6247	2.5011	0.0831
256	3.6076	3.9971	2.8322	0.0820

4.1 Synthetic data: 2D Gaussian

We begin with a two-dimensional Gaussian toy setting, where all key quantities have a closed form. Throughout, we take $P^{a_1} = \mathcal{N}(\mathbf{0}, \left[\begin{smallmatrix}10 & 0\\ 0 & 1\end{smallmatrix}\right])$ and $P^{a_2} = \mathcal{N}(\mathbf{0}, \left[\begin{smallmatrix}1 & 0\\ 0 & 10\end{smallmatrix}\right])$. The base distribution P^0 varies across four settings, described below. Results are presented in Table 1.

Factorized conditionals. For the in-distribution setting, we set $P^0 = \mathcal{N}(\mathbf{0}, \left[\begin{smallmatrix}10 & 0\\ 0 & 10\end{smallmatrix}\right])$, yielding (by (11)) the target composition $\mathbf{P}^w = \mathcal{N}(\mathbf{0}, \left[\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right])$. Because this composed distribution lies within the effective training support of P^0, P^{a_1}, and P^{a_2}, naïve denoising is already highly accurate. Consequently, applying FKC yields no improvements, and actively degrades performance for undertrained models. For the OOD configuration, we set $P^0 = \mathcal{N}(\mathbf{0}, \left[\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right])$, yielding $\mathbf{P}^w = \mathcal{N}(\mathbf{0}, \left[\begin{smallmatrix}10 & 0\\ 0 & 10\end{smallmatrix}\right])$. Although naïve denoising achieves a reasonable fit, out-of-distribution estimation error remains problematic. When FKC is applied, this error compounds drastically, quickly diverging away from the correct distribution as particle count increases, as shown in Figure 1 with $K = 16$ particles.

Non-factorized conditionals. For the in-distribution setting, we set $P^0 = \mathcal{N}(\mathbf{0}, \left[\begin{smallmatrix}20 & 0\\ 0 & 20\end{smallmatrix}\right])$, yielding $\mathbf{P}^w = \mathcal{N}\left(\mathbf{0}, \frac{20}{21} I\right)$. Here, naïve denoising consistently yields poor distributional fit, in line with expectations over approximation error. However, FKC successfully leverages well-trained models to correct this approximation error as the particle count increases. In the OOD setting, we set

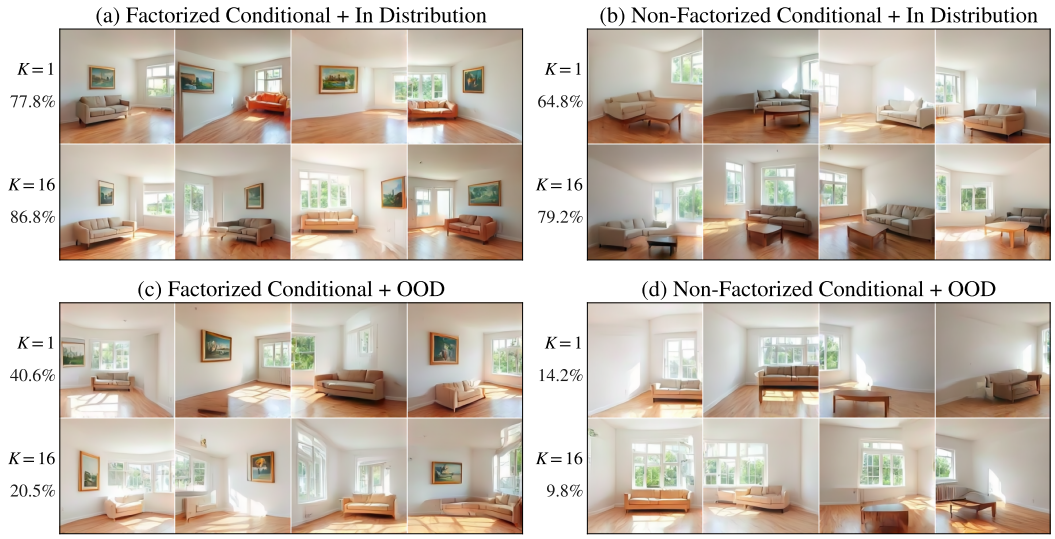


Figure 2: **Couches and paintings (∼factorized) are largely combined accurately, while couches and tables (non-factorized) are not; FKC is only effective when the target composition is ID.** Illustrative samples from base-composed distributions across four experimental settings, comparing naïve denoising ($K = 1$ particles) against FKC ($K = 16$); success rate on joint object presence is evaluated over 500 images per cell. In **(a)** and **(c)**, the target distribution is highly concentrated on empty rooms with exactly one couch *and* one painting; in **(b)** and **(d)**, it is concentrated on rooms with exactly one couch *and* one coffee table. Underlying conditional mixtures are altered to render the target composition ID or OOD. **(a)** Naïve denoising often generates couches *and* paintings; $K = 16$ improves composition slightly by correcting for mild non-factorization-induced approximation error. **(b)** Naïve more frequently fails to generate both a couch *and* table and often warps results; $K = 16$ corrects for the non-factorization. **(c)** Naïve samples may contain both objects but they frequently appear alone; $K = 16$ worsens this as it accumulates estimation error OOD. **(d)** The most challenging setting; couches and tables rarely appear simultaneously and warping is severe.

$P^0 = \mathcal{N}\left(\mathbf{0}, \left[\begin{smallmatrix} 1.1 & 0 \\ 0 & 1.1 \end{smallmatrix}\right]\right)$, so $\mathbf{P}^w = \mathcal{N}\left(\mathbf{0}, \frac{110}{21} I\right)$. In this regime, distributional fit remains poor across all non-analytical settings, and FKC strictly worsens performance when applied to learned models. It is only when the algorithm is supplied with the exact analytical scores that it achieves the expected theoretical improvements. Experiments on Gaussian mixtures are given in App. H.1.

4.2 Image generation: objects in a room and facial features

Semi-synthetic data. To evaluate these dynamics on more realistic, higher-dimensional visual data, we create a semi-synthetic dataset of realistic room images generated via a text-to-image model. The dataset is partitioned into discrete classes: empty rooms, rooms containing a single object, and rooms containing exactly two objects. Each base and conditional distribution is a mixture of images from four underlying classes: "empty," "couch," a second object "X," and "couch + X." We alter object X to either "framed painting" or "coffee table" to control the factorization of the distributions. When X is a framed painting, the conditionals are approximately factorized, as couches and paintings typically occupy disjoint spatial regions in a room (floor versus wall). Conversely, when X is a coffee table, factorization breaks, as both objects compete for overlapping spatial support on the floor.

We define the ID and OOD settings by adjusting the probability weights across the four underlying classes, represented as the vector $(P_{\text{empty}}, P_{\text{couch}}, P_X, P_{\text{couch}+X})$. In the in-distribution setting, P^0 is a uniform mixture $(0.25, 0.25, 0.25, 0.25)$, and the perturbation distributions are heavily biased toward the presence of their respective objects: P^{a_1} is set to $(\alpha, 0.5-\alpha, \alpha, 0.5-\alpha)$ and P^{a_2} is set to $(\alpha, \alpha, 0.5-\alpha, 0.5-\alpha)$, where we use a small smoothing factor $\alpha = 0.01$ to ensure the composition remains well-defined. Under this construction, the composed target distribution places nearly all of its probability mass on the "couch + X" class, which is well-supported within each condition's training data. In the OOD setting, the base distribution is defined as $(1.0 - 3\beta, \beta, \beta, \beta)$ with $\beta = 0.05/3$,

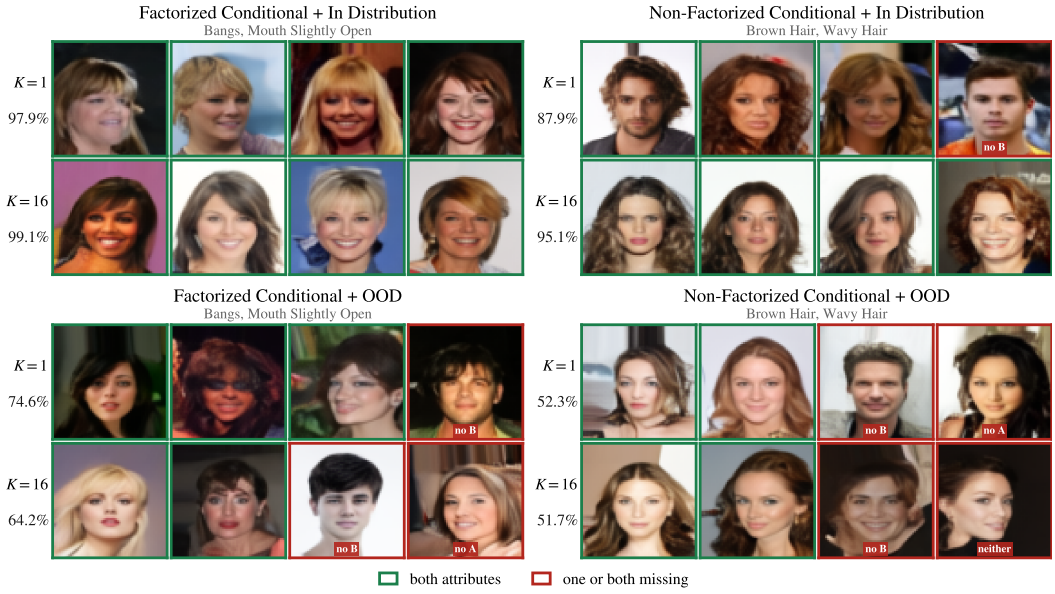


Figure 3: **The same trends also hold for facial image generation**. Setup is similar to Figure 2.

heavily skewing it toward the "empty" class. P^{a_1} and P^{a_2} are identically skewed toward the isolated "couch" and "X" classes, respectively. As a result, the composed target still concentrates almost entirely on the "couch + X" class, which is *not* well-supported within each condition's training data. Illustrative samples from the resulting composed distributions are provided in Figure 2.

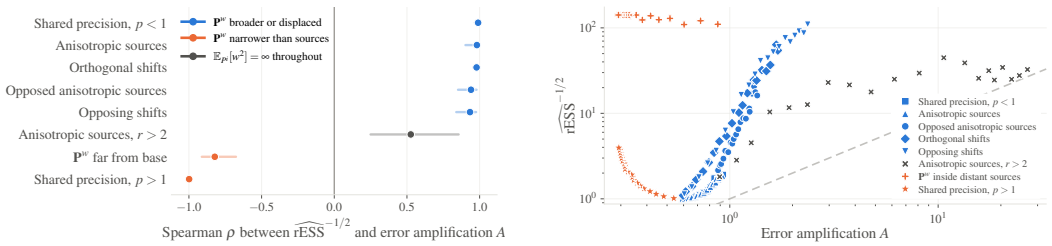
Real-world data. We perform a similar analysis on the CelebA dataset (Liu et al., 2015). Instead of objects, we consider facial attributes, and as a base case we use the distribution over all face images without either facial attribute specified by the perturbations. Our choice of attributes is likewise designed to separate factorized from unfactorized settings: we consider features which are either disjoint ('mouth slightly open' and 'has bangs') or intertwined ('brown hair' and 'wavy hair'). To push the limits of the theory and evaluate compositions which may not even be strictly defined, we set the mixture distributions to $P^0 = (0.99, \alpha, \alpha, \alpha)$ with $\alpha = 1/300$, and set $P^{a_1} = (0, 0.5+\beta, 0, 0.5-\beta)$ and $P^{a_2} = (0, 0, 0.5+\beta, 0.5-\beta)$ where $\beta = 0$ for ID settings and $\beta = 0.5$ for OOD settings. Figure 3 shows that trends similar to the room experiments hold.

4.3 Using κ as a practical sensitivity metric

The sensitivity terms κ_i controlling the upper bound in Theorem 1 have a well-known connection to importance-weighted effective sample size. Specifically, given $X_1, \ldots, X_N \sim P^i$ and weights $w_n \propto P^w(X_n)/P^i(X_n)$, define relative ESS by $\widehat{\mathrm{rESS}}_i = \frac{\left(\sum_{n=1}^N w_n\right)^2}{N \sum_{n=1}^N w_n^2}$. It can be shown that in the population setting, $\kappa_i = \mathrm{rESS}_i^{-1/2}$ (App. G.6). This motivates the question: can we use an estimate of this sensitivity term to determine when compositions are likely to be unstable or unreliable, without requiring ground truth samples from that composition?

To answer this question, we begin with Gaussian experiments where quantities are known in closed form, and we can use the maximum-likelihood estimator as parametric estimators of the source distributions. Figure 4 shows how κ_i tracks error amplification on base compositions of MLE models trained on 20 000 samples per source distribution, across various 2D Gaussian families. To obtain a single summary metric for sensitivity, we aggregate via $\mathrm{rESS} = \min_{i \in \{0, a_1, a_2\}} \mathrm{rESS}_i$, which induces a valid (though slightly looser) sensitivity bound.

The Gaussian setting allows for parametric models, so rESS could be estimated directly. For diffusion models, this is not so simple. However, as it turns out, we can write rESS for a base composition in terms of only two density ratios P^{a_1}/P^0 and P^{a_2}/P^0. Then, adapting a technique from Li et al. (2023), we can approximate these density ratios directly with the trained conditional diffusion model



(a) Within-family ranking of error amplification.

(b) Absolute prediction of error amplification.

Figure 4: **Relative ESS is a strong ranker of within-family error amplification despite at times being a loose bound, but has two distinct failure cases.** Within a single family of Gaussian distributions, $\widehat{\mathrm{rESS}}^{-1/2}$ achieves near perfect correlation with measured error amplification $A = \mathrm{TV}(\mathbf{P}^w, \widehat{\mathbf{P}}^w)/\sum_i \mathrm{TV}(P^i, \widehat{P}^i)$, with two types of exceptions. First, when $\mathbf{P}^w$ is narrower than every source (orange points): the sources cover $\mathbf{P}^w$ poorly so κ rises, but parametric Gaussian fitting errors are smooth and so nearly constant across narrow $\mathbf{P}^w$, and the constant part cancels upon renormalization, so A falls. The second failure mode is when $\mathbb{E}_{P^i}[w^2] = \infty$ (gray points). Here, $\kappa = \infty$, therefore $\widehat{\mathrm{rESS}}$ is determined by the sample's largest weights and ceases to track amplification, shown in (b) by the gray points noisily plateauing.

by performing paired single-step denoising inferences on noised samples from the source distributions. Proofs and further details are given in Apps. G.7 and G.8. Applying this to the image generation experiments, we find that density estimation is too unstable in high-dimensional pixel space, so we use latent diffusion models instead. Although error amplification cannot be measured as ground truth densities are not known, we observe that the metric separates ID from OOD compositions well. In particular, $\log \widehat{\mathrm{rESS}}$ separates OOD compositions from ID compositions by 3.13–3.87 for rooms and 1.07–2.32 for faces, while individual $\log \widehat{\mathrm{rESS}}$ magnitudes range from 2.27–6.93. Further details and experimental results are given in App. H.5.

5 Discussion

In this work, we investigated the challenges of compositional generation, separating the ill-conditioning of the composition operator, distributional estimation error, and inference-time approximation error, and highlighted their interplay.

Key takeaways. Our main upper bound and the matching lower bound (Theorem 1 and Proposition 5) show that the sensitivity of compositional generation is controlled by the χ^2-distance between the target distribution $\mathbf{P}^w$ and the source distributions $\mathbf{P}$: when $\mathbf{P}^w$ is less covered by the sources, compositional generation can become more sensitive to small estimation errors. This characterization can provide crucial guidance for many downstream questions, e.g., when such extrapolations can be trusted (as considered in Sec. 4.3), and how much training data needs to be collected if compositional generation is an important downstream task. Building on this, we highlight a kind of "bias-variance" tradeoff for compositional generation: in certain settings of interest, methods for correcting inference-time approximation error amplify distributional estimation error, motivating the development of inference-time composition methods that adaptively balance these two sources of error.

Limitations. While our primary sensitivity bound is tight, it is a worst-case bound, and experiments demonstrate that it can be loose in practice. It is therefore of interest to understand under what additional, reasonable assumptions compositional error bounds of this form can be improved. In particular, stronger assumptions on the relationship between the source distributions and the target composition, assumptions on the form of the distributional estimation errors motivated by model architecture choices, and bounds specialized to regions of interest, for example by ignoring vanishingly improbable tails which can substantially inflate the worst-case bound, may offer further perspective while remaining relevant to practical settings. Finally, our experiments only considered base-composed distributions, and a more in-depth study across different weightings may provide deeper insights.

AI use statement

In this work, we used generative AI tools to generate a synthetic data set for the room experiments (details of which are in App. I.3), to assist in formulating mathematical claims and writing and simplifying proofs, and to judge images produced by diffusion models for object presence and quality. Additionally, some of the proofs are primarily AI-generated, largely corollaries and examples. We additionally used generative AI tools to implement portions of the experimental code, to programatically modify figures, and to edit text for readability and concision. We have not used generative AI tools to help develop conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments (beyond debugging and code synthesis), to assist with translation, or to clean and reformat datasets.

AI-generated code was tested for correctness, AI-generated proof steps and proof rewrites were checked for correctness, and the AI-generated dataset underwent manual quality filtering. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

All experimental details, including dataset construction, are provided in App. I, along with a link to an anonymous codebase containing experimental code. Proofs for all mathematical claims are provided in App. G.

References

Abhinav Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Chiara Ricci-Tam, Chris Carpenter, Vishvak Subramanyam, Aidan Winters, Mingze Dong, Sravya Tirukkovalur, Jeremy Sullivan, Brian Plosky, Basak Eraslan, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H Roohani. Predicting cellular responses to perturbation across diverse contexts with state. In *NeurIPS 2025 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences*, 2025.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, pp. 372–407. PMLR, 2023.

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.

Arwen Bradley and Preetum Nakkiran. Classifier-free guidance is a predictor-corrector. *arXiv preprint arXiv:2408.09000*, 2024.

Arwen Bradley, Preetum Nakkiran, David Berthelot, James Thornton, and Joshua M. Susskind. Mechanisms of projective composition of diffusion models. In *Forty-second International Conference on Machine Learning*, 2025.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *Advances in Neural Information Processing Systems*, 36:45419–45462, 2023.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B. Burkhardt, Andrea Califano, Jonah Cool, Abby F. Dernburg, Kirsty Ewing, Emily B. Fox, Matthias Haury, Amy E. Herr, Eric Horvitz, Patrick D. Hsu, Viren Jain, Gregory R. Johnson, Thomas Kalil, David R. Kelley, Shana O. Kelley, Anna Kreshuk, Tim Mitchison, Stephani Otte, Jay Shendure, Nicholas J. Sofroniew, Fabian Theis, Christina V. Theodoris, Srigokul Upadhyayula, Marc Valer, Bo Wang, Eric Xing, Serena Yeung-Levy, Marinka Zitnik, Theofanis Karaletsos, Aviv Regev, Emma Lundberg, Jure Leskovec, and Stephen R. Quake. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, Dec 2024. ISSN 0092-8674.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Out-of-distribution generalization in kernel regression. *Advances in Neural Information Processing Systems*, 34:12600–12612, 2021.

Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 129–136. JMLR Workshop and Conference Proceedings, 2010.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

R. Douc and O. Cappe. Comparison of resampling schemes for particle filtering. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, pp. 64–69, 2005. doi: 10.1109/ISPA.2005.195385.

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*, pp. 8489–8510. PMLR, 2023.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Rémi Flamary, Cédric Vincent-Cuaz, Nicolas Courty, Alexandre Gramfort, Oleksii Kachaiev, Huy Quang Tran, Laurène David, Clément Bonet, Nathan Cassereau, Théo Gnassounou, Eloi Tanguy, Julie Delon, Antoine Collas, Sonia Mazelet, Laetitia Chapel, Tanguy Kerdoncuff, Xizheng Yu, Matthew Feickert, Paul Krzakala, Tianlin Liu, and Eduardo Fernandes Montesuma. Pot python optimal transport (version 0.9.5), 2024. URL https://github.com/PythonOT/POT.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Samory Kpotufe and Guillaume Martinet. Marginal singularity, and the benefits of labels in covariate-shift. In *Conference On Learning Theory*, pp. 1882–1886. PMLR, 2018.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025.

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2206–2217. IEEE, 2023.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5404–5411, 2024.

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

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*, pp. 423–439. Springer, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L. Ibarra, Sanjay R. Srivatsan, Mohsen Naghipourfar, Riza M. Daza, Beth Martin, Erez Lieberman Aiden, Jay Shendure, José L. McFaline-Figueroa, Pierre Boyeau, and Fabian J. Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023.

Jens Müller, Robert Schmier, Lynton Ardizzone, Carsten Rother, and Ullrich Köthe. Learning robust models using the principle of independent causal mechanisms. In *DAGM German Conference on Pattern Recognition*, pp. 79–110. Springer, 2021.

Kartik Narayan, Vibashan VS, Rama Chellappa, and Vishal M Patel. Facexformer: A unified transformer for facial analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11369–11382, 2025.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Emmanuel Noutahi, Jason Hartford, Prudencio Tossou, Shawn Whitfield, Alisandra K. Denton, Cas Wognum, Kristina Ulicna, Michael Craig, Jonathan Hsu, Michael Cuccarese, Emmanuel Bengio, Dominique Beaini, Christopher Gibson, Daniel Cohen, and Berton Earnshaw. Virtual cells: Predict, explain, discover. *arXiv preprint arXiv:2505.14613*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Featured Certification.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Yinuo Ren, Wenhao Gao, Lexing Ying, Grant M. Rotskoff, and Jiequn Han. Driftlite: Lightweight drift control for inference-time scaling of diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42:927–935, 2024.

D. Sanz-Alonso S. Agapiou, O. Papaspiliopoulos and A. M. Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, pp. 405–431, 2017.

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

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

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 *Forty-second International Conference on Machine Learning*, 2025.

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

Chandler Squires and Caroline Uhler. Causal structure learning: A combinatorial perspective. *Foundations of Computational Mathematics*, 23(5):1781–1815, 2023.

Chandler Squires, Anna Seigal, Salil S Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning*, pp. 32540–32560. PMLR, 2023.

Simo Särkkä and Arno Solin. *Applied Stochastic Differential Equations*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2019.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2314–2322. PMLR, 2024.

Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. Linear causal representation learning from unknown multi-node interventions. *Advances in Neural Information Processing Systems*, 2024.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning: Linear and general transformations. *Journal of Machine Learning Research*, 26(112):1–90, 2025.

Burak Varıcı, Chandler Squires, and Pradeep Ravikumar. Causal representation learning. *Neurosymbolic AI: Foundations and Applications*, pp. 307–346, 2026.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Gefei Wang, Tianyu Liu, Jia Zhao, Youshu Cheng, and Hongyu Zhao. Modeling and predicting single-cell multi-gene perturbation responses with sclambda. *bioRxiv*, 2024.

Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. Concept algebra for (score-based) text-controlled generative models. *Advances in Neural Information Processing Systems*, 2023.

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Liam Hodgson, Ihab Bendidi, Cian Eastwood, Shawn T. Whitfield, Craig Russell, Yassir El Mesbahi, Jiarui Ding, Marta M. Fay, Berton Earnshaw, Emmanuel Noutahi, and Alisandra K. Denton. Txpert: using multiple knowledge graphs for prediction of transcriptomic perturbation effects. *Nature Biotechnology*, 2026.

Yu Xie, Ludwig Winkler, Lixin Sun, Sarah Lewis, Adam Foster, Jose Jimenez-Luna, Tim Hempel, Michael Gastegger, Yaoyi Chen, Iryna Zaporozhets, Cecilia Clementi, Christopher M. Bishop, and Frank Noe. Enhanced diffusion sampling: Efficient rare event sampling and free energy calculation with diffusion models. In *ICML 2026 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2026.

Zuheng Xu, Moksh Jain, Ali Denton, Shawn Whitfield, Aniket Didolkar, Berton Earnshaw, and Jason Hartford. Automated discovery of pairwise interactions from unstructured data, 2024.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 36:50254–50292, 2023.

Contents of Appendix

A Formalization of density-related assumptions **16**

B Weighted composition: A causal representation learning perspective **16**

C Diffusion model background **17**

C.1 Feynman-Kac correctors 18

D Quantifying out-of-distribution support mismatch **19**

E Further detail on factorized conditionals **20**

E.1 Relaxations 20

E.2 Quantifying approximate factorization in real-world settings 21

F Base-composition versus Bayes composition **21**

G Proofs and additional theoretical details **23**

G.1 Preliminaries: Gaussian weighted composition 23

G.2 Arbitrary instability under vanishing source error 24

G.3 Local sensitivity to relative L^2 source error 26

G.4 Inference-time approximation error in Gaussians 27

G.5 Feynman-Kac correctors under factorized conditionals 30

G.6 Connection of Theorem 1 to importance-sampling effective sample size 31

G.7 Importance ratios for a base composition 32

G.8 Estimating ESS ratios with a conditional diffusion model 32

G.9 Global stability under stronger pathwise assumptions 33

G.10 A direct global bound from combined relative L^2 error 34

G.11 TV bound through uniform control of log-density error 35
G.12 Stability conditions in Bayes composition 35

H Additional experimental results **38**
H.1 Synthetic data: Gaussian mixture models 38
H.2 Image generation: full accuracies and warping scores 38
H.3 Latent diffusion models 39
H.4 Separate diffusion models 40
H.5 Sensitivity metric in image generation 41

I Experiment setup details **43**
I.1 2D Gaussian 43
I.2 Mixture of Gaussians 44
I.3 Objects in a room 44
I.4 Facial image generation 46
I.5 Latent diffusion models 47
I.6 Sensitivity metric experiments 47

J Additional room images **49**

K Tables **50**
K.1 Two-dimensional Gaussian (conditional model) 50
K.2 Two-dimensional Gaussian (separate models) 52
K.3 Gaussian mixture models (conditional model) 54
K.4 Gaussian mixture models (separate models) 55

A Formalization of density-related assumptions

Our density-related assumptions in Sec. 2 can be more formally stated as follows:

Assumption 1. *There exists some base measure μ on $\mathcal{X}$, with* $\text{supp}(\mu) = \mathcal{X}$*, such that $P_a \ll \mu$ and $\mu \ll P_a$ for all $a \in \mathcal{A}_o$, where $\ll$ denotes the absolute continuity relation.*

First, by the Radon-Nikodym theorem, the condition that $P^a << \mu$ ensures that each distribution P^a has a density with respect to μ, and that these densities are uniquely defined μ-almost everywhere. Second, this condition implies that $P^a(x) > 0$ for μ-almost all $x \in \mathcal{X}$. Hence, our equality statements should be formally interpretated as holding over this *equivalence class* of densities.

B Weighted composition: A causal representation learning perspective

In causal representation learning, one often considers *causal representation models* as an inductive bias for compositional generalization. In particular, one can define a ***causal representation model on*** $\mathcal{X}$ as a tuple $M = (S, g)$, where S is a *structural causal model* on $\mathcal{Z} = \mathbb{R}^d$, and $g \colon \mathcal{Z} \to \mathcal{X}$ is a diffeomorphism onto its image (see Varıcı et al. (2026, Definition 10.7)). Most importantly for our purposes, the ***latent-space observational distribution*** of $M = (S, g)$ has the form

$$Q^{\text{obs}}(z) := \prod_{i=1}^{d} Q^{\text{obs}}(z_i \mid \text{pa}_{\mathcal{G}}(z_i)),$$

where $\mathcal{G}$ is the *causal DAG* associated with the structural causal model S, and the ***feature-space observational distribution*** of M is

$$P^{\text{obs}}(x) = Q^{\text{obs}}(g^{-1}(x)) \cdot |\det J_{g^{-1}}(x)|.$$

Then, an ***intervention*** I on M, with ***targets*** $T(I) \subseteq [d]$, is a collection of conditional distributions $(Q^I(z_i \mid \text{pa}_{\mathcal{G}}(z_i))_{j \in T(I)}$, giving the ***latent-space interventional distribution***

$$Q^I(z) := \prod_{i=1}^{d} Q^{\text{obs}}(z_i \mid \text{pa}_{\mathcal{G}}(z_i))^{\mathbb{1}_{i \notin T(I)}} \cdot Q^I(z_i \mid \text{pa}_{\mathcal{G}}(z_i))^{\mathbb{1}_{i \in T(I)}},$$

and the ***feature-space interventional distribution*** $P^I(x) = Q^I(g^{-1}(x)) \cdot |\det J_{g^{-1}}(x)|$. From these definitions, we see that

$$\frac{Q^I(z)}{Q^{\text{obs}}(z)} = \prod_{i \in T(I)} \frac{Q^I(z_i \mid \text{pa}_{\mathcal{G}}(z_i))}{Q^{\text{obs}}(z_i \mid \text{pa}_{\mathcal{G}}(z_i))},$$

and, since the Jacobian term cancels out,

$$\frac{P^I(x)}{P^{\text{obs}}(x)} = \prod_{i \in T(I)} \frac{P^I(z_i \mid \text{pa}_{\mathcal{G}}(z_i))}{P^{\text{obs}}(z_i \mid \text{pa}_{\mathcal{G}}(z_i))}, \quad \text{where } z = g^{-1}(x). \tag{5}$$

Perturbations as unknown-target interventions In *interventional causal representation learning*, one typically assumes that each *single perturbation* $a = \mathbf{e}_k$ for $k \in [K]$ corresponds to an intervention I_a with unknown targets (Varıcı et al., 2024), though possibly with restrictions on the size of the intervention (Ahuja et al., 2023; Squires et al., 2023; Buchholz et al., 2023; Zhang et al., 2023; Varici et al., 2024; 2025). In particular, to learn the model M from such data, one must also learn a map $\rho \colon a \mapsto I_a$, mapping each perturbation to its representation as an intervention.

Representing *double perturbations* as vectors $\mathbf{e}_{kk'} := \mathbf{e}_k + \mathbf{e}_{k'}$ for $k \neq k'$, (a form of) the principle of *independent causal mechanisms (ICM)* states that the map ρ has a certain modularity property (Müller et al., 2021). In particular, if $\mathbf{e}_k$ and $\mathbf{e}_{k'}$ affect different causal mechanisms (i.e., $T(I_{\mathbf{e}_k}) \cap T(I_{\mathbf{e}_{k'}}) = \varnothing$), then the ICM principle encourages one to assume that $\rho(\mathbf{e}_{kk'}) = I_{\mathbf{e}_k} \cup I_{\mathbf{e}_{k'}}$, i.e., that the corresponding intervention $I_{\mathbf{e}_{kk'}}$ *combines* the interventions $I_{\mathbf{e}_k} = (Q^{I_{\mathbf{e}_k}}(z_i \mid \text{pa}_{\mathcal{G}}(z_i))_{j \in T(I_{\mathbf{e}_k})}$ and $I_{\mathbf{e}_{k'}} = Q^{I_{\mathbf{e}_{k'}}}(z_i \mid \text{pa}_{\mathcal{G}}(z_i))_{j \in T(I_{\mathbf{e}_{k'}})}$, without any changes to the interventional mechanisms.

Weighted composition For consistency with Section 2, let $P^a := P^{I_a}$. By Equation (5) and the principle of independent mechanisms, we obtain

$$\frac{P^{\mathbf{e}_{kk'}}(x)}{P^{\text{obs}}(x)} = \prod_{i \in T(I_{\mathbf{e}_{kk'}})} \frac{P^{\mathbf{e}_{kk'}}(z_i \mid \text{pa}_{\mathcal{G}}(z_i))}{P^{\text{obs}}(z_i \mid \text{pa}_{\mathcal{G}}(z_i))}, \qquad \text{(Equation (5))}$$

$$= \left(\prod_{i \in T(I_{\mathbf{e}_k})} \frac{P^{\mathbf{e}_k}(z_i \mid \text{pa}_{\mathcal{G}}(z_i))}{P^{\text{obs}}(z_i \mid \text{pa}_{\mathcal{G}}(z_i))} \right) \cdot \left(\prod_{i \in T(I_{\mathbf{e}_{k'}})} \frac{P^{\mathbf{e}_{k'}}(z_i \mid \text{pa}_{\mathcal{G}}(z_i))}{P^{\text{obs}}(z_i \mid \text{pa}_{\mathcal{G}}(z_i))} \right) \qquad \text{(ICM)}$$

$$= \frac{P^{\mathbf{e}_k}(x)}{P^{\text{obs}}(x)} \cdot \frac{P^{\mathbf{e}_{k'}}(x)}{P^{\text{obs}}(x)}, \qquad \text{(Equation (5))}$$

where we have used $z = g^{-1}(x)$ throughout. Thus, rearranging and assuming $P^{\mathbf{0}} = P^{\text{obs}}$, we obtain the desired composition: $P^{\mathbf{e}_{kk'}}(x) = P^{\mathbf{e}_k}(x) \cdot P^{\mathbf{e}_{k'}}(x) \cdot (P^{\mathbf{0}}(x))^{-1}$. This argument naturally extends to combinations of *several* perturbations with non-overlapping intervention targets.

Notably, such interventional distributions are generalizations of the *Factorized Conditionals* condition from (Bradley et al., 2025): the assumption of mutually independent sets of variables corresponds to *disconnected* groups of nodes in a causal graph, giving their Equation (7) as a special case where each intervention targets one of these disconnected groups of nodes.

C Diffusion model background

In this section, we review relevant background on conditional diffusion models, using the convention that $t = 0$ is the data distribution and $t = 1$ is the noised distribution, with W_t representing the standard Wiener process and $\overline{W}_t$ denoting the reverse-time Wiener process. Throughout, we consider fixed schedules $(\mu_t)_t$ and $(\sigma_t)_t$, and we write $u_t(x) := \mu_t x$.

Noising First, we defines a *noising operator* $\text{Noise}_t \colon \mathcal{X} \to \mathcal{P}(\mathcal{X})$, which maps x_0 to the marginal distribution of $\mathbf{X}_t$ after Itó integration of the *noising SDE* $\mathrm{d}x_t = u_t(x_t)\mathrm{d}t + \sigma_t \mathrm{d}W_t$.[5] Extending this operator to distributions, we have $\text{Noise}_t \colon \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$ where $\text{Noise}_t \colon P \mapsto \int_{\mathcal{X}} \text{Noise}_t(x) P(\mathrm{d}x)$. As shorthand, given $P^a \in \mathcal{P}(\mathcal{X})$, we write $P^a_t := \text{Noise}_t(P^a)$. Similar, we write $S^a_t(x) := \nabla_{\mathsf{x}} \log P^a_t(x)$ for the Stein score function of P^a_t. Finally, for any noising SDE, there exists a corresponding deterministic ODE with identical marginal distributions (Song et al., 2021). Hence, we define this *noising ODE* as $\mathrm{d}x_t = \tilde{u}_t(x_t)\mathrm{d}t$ for $\tilde{u}_t(x) := u_t(x) - \frac{1}{2}\sigma_t^2 S^a_t(x)$.

Denoising Diffusion models leverage the fact that the stochastic process $(\mathbf{X}^a_t)_t$ also solves the *denoising SDE*

$$\mathrm{d}x^a_t = v^a_t(x^a_t)\mathrm{d}t + \sigma_t \mathrm{d}\overline{W}_t, \quad \text{for} \quad v^a_t(x) := u_t(x) - \sigma_t^2 S^a_t(x), \tag{6}$$

as given in Song et al. (2021). Alternatively, the *denoising ODE* is defined by running the noising ODE backward in time. In practice, the conditional score functions S^a_t can be estimated by a variety of methods (Hyvärinen, 2005; Vincent, 2011; Ho et al., 2020; Song et al., 2021; Lipman et al., 2023; Albergo et al., 2025), with the conditional score functions typically parameterized as conditional deep networks (Perez et al., 2018; Ho & Salimans, 2022); our experiments use denoising diffusion (Ho et al., 2020). Hence, each $P^a_\star$ is estimated by a distribution $\widehat{P}^a$, implicitly defined as the marginal distribution of $\mathbf{X}_0$, after running the denoising SDE in Equation (6) with $\widehat{S}^a(\cdot; t)$.

Composition Given a target distribution $\mathbf{P}^w := \text{Comp}_w(\mathbf{P})$, we could obtain samples from $\mathbf{P}^w$ by running Equation (6) with $S^w_t(x) := \nabla_{\mathsf{x}} \log \mathbf{P}^w_t(x)$ for $\mathbf{P}^w_t := \text{Noise}_t(\mathbf{P}^w)$. However, we do not know $S^w_t(x)$ or have a simple way to compute it. Instead, a simple heuristic used in many prior works (e.g., classifier-free guidance (Ho & Salimans, 2022) and compositional visual generation (Liu et al., 2022)) substitutes this score with a 'naive' additive heuristic, $S^{w,\text{nv}}_t(x) := \sum_{a \in \mathcal{A}_\circ} w_a S^a_t(x)$.

[5] In particular, $\text{Noise}_t(x_0) = \mathcal{N}(\alpha_t x_0, \gamma_t^2 I)$ for $\alpha_t := \exp\left(\int_0^t \mu_s \mathrm{d}s\right)$ and $\gamma_t := \alpha_t \sqrt{\int_0^t \frac{\sigma_s^2}{\alpha_s^2} \mathrm{d}s}$ (Albergo et al., 2025).

The corresponding ***naïve denoising SDE*** is thus given by

$$\mathrm{d}x_t^{w,\mathrm{nvs}} = v_t^{w,\mathrm{nv}}(x_t^{w,\mathrm{nvs}})\mathrm{d}t + \sigma_t \mathrm{d}\overline{W}_t, \quad \text{where} \quad v_t^{w,\mathrm{nv}}(x) := u_t(x) - \sigma_t^2 S_t^{w,\mathrm{nv}}(x), \tag{7}$$

and the corresponding ***naïve denoising ODE*** is

$$\mathrm{d}x_t^{w,\mathrm{nvo}} = \tilde{v}_t^{w,\mathrm{nv}}(x_t^{w,\mathrm{nvo}})\mathrm{d}t \quad \text{where} \quad \tilde{v}_t^{w,\mathrm{nv}}(x) := u_t(x) - \frac{1}{2}\sigma_t^2 S_t^{w,\mathrm{nv}}(x). \tag{8}$$

Notably, the definition of weighted composition ensures that $S_t^{w,\mathrm{nv}}(x) = S_t^w(x)$ at $t = 0$, and the definition of the noising SDE ensures that, if the terminal noising distribution is condition-independent and $\sum_a w_a = 1$, then $S_t^{w,\mathrm{nv}}(x) = S_t^w(x)$ at $t = 1$. However, it is not generically true that $S_t^{w,\mathrm{nv}}(x) = S_t^w(x)$ for intermediate times $t \in (0, 1)$, as noted in several prior works, e.g. Bradley & Nakkiran (2024); Skreta et al. (2025). Hence, naïve denoising methods can introduce substantial inference-time approximation error, as we illustrate in App. G.4.

C.1 Feynman-Kac correctors

In this section, we present the theoretical intuition behind Feynman-Kac Correctors (FKC) (Skreta et al., 2025). At the theoretical level, Skreta et al. (2025) derive the importance weights required to correct the inference-time dynamics of a denoising diffusion process toward the desired intermediate distributions, termed Feynman-Kac Correctors. To use these weights in practice, they propose a Sequential Monte Carlo (SMC) scheme as a practical instantiation for realizing this correction with finitely many particles.

Although the naïve denoising SDE uses a *score* that matches S_t^w at $t = 0$, its marginal at $t = 0$ is not $\mathbf{P}^w$: the particle position at the endpoint depends on integrated dynamics over the entire trajectory, and $S_t^{w,\mathrm{nv}}$ is not correct at $t > 0$. FKC addresses this inference-time approximation error by maintaining K weighted particles $\{(x_t^{(k)}, \omega_t^{(k)})\}_{k=1}^K$, with $\omega_t^{(k)}$ the log importance weight of particle k, such that the weighted empirical *distribution* tracks a target along the entire denoising trajectory.

The natural target $\mathbf{P}_t^w = \mathrm{Noise}_t(\mathbf{P}^w)$ has an intractable score S_t^w, so FKC tracks instead the weighted composition of the noised marginals, $\mathrm{Comp}_w(\{P_t^a\}_a)$, whose score is the naïvely-composed $S_t^{w,\mathrm{nv}}$. This proxy agrees with $\mathbf{P}_t^w$ at $t = 0$ (Noise_0 is the identity) but diverges from it at intermediate t, since noising and composition do not commute. Sampling correctness thus holds, at $t = 0$ the proxy equals $\mathbf{P}^w$ in distribution, so the weighted particles approximate $\mathbf{P}^w$.

Determining the weight update to achieve this follows from comparing how the proposal $\widetilde{P}_t^{w,\mathrm{nv}}$ and the proxy $\mathrm{Comp}_w(\{P_t^a\}_a)$ evolve in time. Both satisfy Fokker-Planck-type equations involving the drift divergence, score, and Laplacian of the log density. Subtracting one from the other, they leave a residual g_t that the log-weights along a particle trajectory must account for:

$$\mathrm{d}\omega_t(x_t) \;=\; \bar{g}_t(x_t)\,\mathrm{d}t, \qquad \bar{g}_t(x) \;:=\; g_t(x) - \mathbb{E}_{X_t \sim \mathrm{Comp}_w(\{P_t^a\}_a)}[g_t(X_t)], \tag{9}$$

where the centering enforces $\int \bar{g}_t(x)\, P_t^w(x)\,\mathrm{d}x = 0$, preventing uniform divergence or collapse of all log-weights simultaneously. For weighted compositions, g_t is given in closed form by:

$$g_t(x) \;=\; \Big(1 - \sum\nolimits_{a \in \mathcal{A}_o} w_a\Big)\,\langle \nabla_{\mathsf{x}}, u_t(x)\rangle + \frac{\sigma_t^2}{2}\Big(\sum\nolimits_{a \in \mathcal{A}_o} w_a\, \|S_t^a(x)\|^2 - \|S_t^{w,\mathrm{nv}}(x)\|^2\Big). \tag{10}$$

To utilize these weights $\omega_t^{(k)}$ to perform corrections, a Sequential Monte Carlo method is employed. During denoising, resampling is performed at each step based on the increment $\omega_t^{(k)} = g_t(x_t^{(k)})$ via systematic resampling proportional to $\exp\left(\omega_t^{(k)}\right)$ (Douc & Cappe, 2005). Note that when $K = 1$, the ensemble consists of only a single particle, so resampling has no effect and running FKC is equivalent to running the naïve denoising SDE.

We also note how as an additional nicety of base-composed distributions (2), the divergence term in (10) cancels to 0 and can be ignored even for non-linear drifts. (For linear drifts, it is constant and can be ignored regardless).

The FKC framework extends naturally to denoising ODEs, with the same g_t as in the SDE case. However, applying the instantiation of the FKC framework as described with SMC rejection sampling is less well-motivated for ODEs because their trajectories are deterministic. Because the denoising ODE lacks noise injection from the Wiener process, particle trajectories never stochastically diverge. When FKC resampling duplicates a high-weight particle, those identical particles will follow the exact same deterministic trajectory for the remainder of the reverse process. This means that resampling monotonically decreases the diversity of the particle swarm. Despite this limitation, if the continuity-equation residual is large, FKC can remain mathematically helpful to route the probability mass to the correct target coordinates.

D Quantifying out-of-distribution support mismatch

Our definition of out-of-distribution composition primarily concerns whether samples from the target composition P^w occupy low-density regions under the source distributions. To quantify this notion directly, for each source distribution P^a, let $H_a^{0.95}$ denote its 95%-mass highest-density region (HDR), and define

$$L_a := \mathbb{P}_{x\sim P^w}\left[x \notin H_a^{0.95}\right].$$

Thus, L_a measures the fraction of the target distribution that lies outside the 95% highest-density region of source P^a. Since our experiments use three source distributions P^0, P^{a_1}, and P^{a_2}, we additionally report the fraction of target mass lying outside the 95% HDRs of all three sources,

$$L_{\text{all}} := \mathbb{P}_{x\sim P^w}\left[x \notin H_0^{0.95},\ x \notin H_{a_1}^{0.95},\ x \notin H_{a_2}^{0.95}\right].$$

We compute L_a and L_{all} for the synthetic settings using the ground-truth distributions. For the two-dimensional Gaussian experiments, the target composition is a Gaussian whose precision is given in closed form by (11). The 95% HDR of each source is the ellipsoid $H_a^{0.95} = \{x : x^\top \Sigma_a^{-1} x \le \chi^2_{2,0.95}\}$. We draw $n = 10^6$ exact samples $x_i \sim P^w$ and estimate $L_a = \frac{1}{n}\sum_i \mathbf{1}[x_i^\top \Sigma_a^{-1} x_i > \chi^2_{2,0.95}]$. L_{all} is estimated the same way, counting a sample only if it falls outside all three ellipsoids at once.

For the Gaussian mixture experiments, the HDRs have no closed form, so we estimate each one. For each source P^a, we draw $m = 10^5$ calibration samples $y_1, \ldots, y_m \sim P^a$, evaluate the exact log density $\log p^a(y_k)$, and set τ_a to the 5th percentile of these values. This gives $H_a^{0.95} \approx \{x : \log p^a(x) \ge \tau_a\}$ and $L_a \approx \frac{1}{n}\sum_i \mathbf{1}[\log p^a(x_i) < \tau_a]$. Because $P^w \propto P^{a_1}P^{a_2}/P^0$ is not itself a gaussian mixture, we draw $n = 5{,}000$ target samples from the exact ratio density. To do so, we use rejection sampling in the ID setting and importance sampling with $1{,}000\times$ oversampling in the OOD setting.

Table 2: **Target support mismatch under the source distributions.** Each source column reports L_a, the probability that a target sample lies outside the 95% highest-density region of that source. The final column reports L_{all}, the probability that a target sample lies outside the 95% highest-density regions of all three sources simultaneously.

Dataset	Setting	Outside $H_0^{0.95}$	Outside $H_{a_1}^{0.95}$	Outside $H_{a_2}^{0.95}$	All three
Gaussian	FC + ID	$< 0.01\%$	1.5%	1.5%	$< 0.01\%$
Gaussian	NFC + ID	$< 0.01\%$	1.3%	1.3%	$< 0.01\%$
Gaussian	FC + OOD	74.1%	48.9%	48.9%	28.5%
Gaussian	NFC + OOD	53.3%	31.0%	31.0%	12.2%
GMM	ID	5.1%	5.1%	5.3%	4.8%
GMM	OOD	95.3%	97.7%	97.6%	95.3%

Table 2 confirms the intended distinction between the ID and OOD settings. In the ID settings, the targets remain predominantly within high-density regions of the sources, whereas in the OOD settings the targets place substantial mass in regions that are low-density under every source.

For high-dimensional settings such as the image generation experiments, quantifying support mismatch is more difficult, and many notions of distributional distance or density are strongly dependent on the chosen representation or featurizer. Nevertheless, our experimental design makes it clear whether the target class is represented in the source distributions. For instance, in the room experiments, the target composition is concentrated on images containing both a couch and a second object X, where

X is either a framed painting or a coffee table. In the ID setting, images containing both objects have substantial probability under the source distributions. In the OOD setting, each source assigns only small probability to the "couch + X" class, while the composed target is concentrated almost entirely on this class, so the composed distribution is poorly supported by each of its source distributions.

E Further detail on factorized conditionals

E.1 Relaxations

Factorized Conditionals can be relaxed in two important ways. First, factorization may hold after an invertible reparameterization into a feature space, so factorized components may correspond to concepts rather than, e.g., blocks of pixels. Second, the independence assumptions can be relaxed to approximate versions, in which case a bound on the resulting composition is obtained.

Factorization in a feature space. Suppose there exists a C^1 diffeomorphism $A : \mathbb{R}^n \to \mathbb{R}^n$ such that the transformed distributions

$$A_{\#}P^0,\ A_{\#}P^{a_1}, \ldots, A_{\#}P^{a_k}$$

satisfy Factorized Conditionals with respect to a partition $M_0, M_1, \ldots, M_k$, where $A_{\#}$ denotes the pushforward by A. Bradley et al. (2025) show that, for base compositions, the composition operator is equivariant under such reparameterizations,

$$\mathrm{COMP}_w(\{A_{\#}P^a\}_a) = A_{\#}\,\mathrm{COMP}_w(\{P^a\}_a)\,.$$

As a result, for the base composition P^w, the transformed distribution satisfies

$$A_{\#}P^w(z) = (A_{\#}P^0)(z_{M_0}) \prod_{i=1}^{k} (A_{\#}P^{a_i})(z_{M_i}).$$

That is, the marginal of $A_{\#}P^w$ on each block M_i is exactly the corresponding marginal of $A_{\#}P^{a_i}$. For example, factorized blocks could represent an image's content (e.g. "cat" or "dog") versus its style (e.g. "realistic" or "impressionist"). Moreover, this feature transformation need not be known in order to perform the composition.

However, this does not imply that denoising diffusion sampling is exact for an arbitrary invertible transformation. Bradley et al. (2025) show that if $A(x) = Ax$ for an orthogonal matrix A, then isotropic diffusion noising commutes with the transformation, and naïve denoising using the composed scores exactly samples from P^w. For a general invertible transformation, however, the noising process need not commute with the transformation, so naïve denoising need not sample from the correct distribution.

Approximate factorized conditionals. The exact independence assumptions can also be partially relaxed. We will retain the requirement that the active blocks $M_1, \ldots, M_k$ are mutually independent under the base distribution and that each conditional equals the base distribution outside its active block, but replace exact independence between each block and its complement with approximate independence. In particular, assume

$$\sup_{x_{M_i^c}} D_{\mathrm{KL}}\big(P^{a_i}(x_{M_i}) \,\big\|\, P^{a_i}(x_{M_i} \mid x_{M_i^c})\big) \le \epsilon_i,$$

together with an analogous bound ϵ_0 for the background block M_0.

Let

$$P^\star(x) = P^0(x_{M_0}) \prod_{i=1}^{k} P^{a_i}(x_{M_i})$$

denote the distribution obtained by enforcing exact blockwise factorization while preserving the corresponding block marginals, where P^w denotes the actual base composition. Under these assumptions, Bradley et al. (2025) show that

$$D_{\mathrm{KL}}\left(P^\star \,\|\, P^w\right) \le \epsilon_0 + \sum_{i=1}^{k} \epsilon_i.$$

That is, the consequences of the Factorized Conditional structure are stable to small violations of the exact independence assumptions, we do not require exact Factorized Conditionals for the composed distribution to remain close to the distribution obtained by exactly factorizing the same block marginals. Note this is a distribution-level result concerning the composition operator, it does not provide any guarantees for diffusion sampling unlike the first relaxation.

E.2 Quantifying approximate factorization in real-world settings

It is not possible in general to determine whether a collection of distributions is factorized conditional (or whether a suitable feature space exists such that they factorize within it), or how close a collection of distributions are to being factorized conditional, without access to ground-truth distributions. We therefore report an empirical metric motivated by the observation from Bradley et al. (2025) that, under exact Factorized Conditionals, the mean-effect vectors necessarily satisfy the orthogonality condition

$$(\mu_{a_1} - \mu_0)^\top (\mu_{a_2} - \mu_0) = 0,$$

where $\mu_a = \mathbb{E}_{x \sim P^a}[x]$ is the mean vector of condition a. We measure the unsigned cosine similarity

$$\frac{\left|(\mu_{a_1} - \mu_0)^\top (\mu_{a_2} - \mu_0)\right|}{\|\mu_{a_1} - \mu_0\| \, \|\mu_{a_2} - \mu_0\|},$$

where a value near 0 suggests (but does not guarantee) that the distributions are more FC-like in the chosen representation, while a magnitude near 1 demonstrates that they are not.

To obtain means in a reasonable feature space, we pass each image in our dataset through an encoder before computing this diagnostic. Specifically, for an encoder ϕ, we use $\mu_a = \mathbb{E}_{x \sim P^a}[\phi(x)]$. We report results for CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2024), and a custom VAE trained only on the room dataset or CelebA. These effectively place an upper bound on the factorization heuristic, as a more suitable feature space for this particular setting may exist since two of the encoders are general-purpose, and one is our VAE, which was trained on a dataset small enough that it may not learn disentangled representations.

Table 3: **Unsigned cosine similarity between the mean-effect vectors for the couch-painting and couch-table settings under different feature encoders.** Values closer to zero are more consistent with the orthogonality condition implied by factorized conditionals.

Feature encoder	Couch, painting	Couch, table
CLIP ViT-B/32	0.216	0.431
DINOv2 Large	0.305	0.475
VAE	0.371	0.551

Table 4: **Unsigned cosine similarity between the mean-effect vectors for the CelebA attributes under different feature encoders.** Values closer to zero are more consistent with the orthogonality condition implied by factorized conditionals.

Encoder	Bangs + Mouth Slightly Open (FC)	Brown Hair + Wavy Hair (NFC)
CLIP ViT-B/32	0.048	0.447
DINOv2 Large	0.042	0.634
VAE	0.030	0.793

Table 3 and Table 4 clearly indicate that couches and paintings are *more* factorized than couches and tables, and that the CelebA attributes chosen for FC are *more* factorized than those chosen for NFC. However, this heuristic cannot provide an absolute quantification of factorization on its own.

F Base-composition versus Bayes composition

It is worth being precise about the relationship between the general forms of composition we study, and a particularly common form of composition, which, following the notation of Bradley et al.

(2025), we will refer to as Bayes composition (Liu et al., 2022; Du et al., 2023). Bayes compositions use the unconditional distribution as the denominator (while other weighted compositions such as classifier-free guidance (Ho & Salimans, 2022) similarly use the unconditional distribution as a reference with a negative weight). However, our P^0 does not always denote the unconditional marginal over all conditions. In many cases of interest, it instead denotes a reference or control condition, for example, a control trial, e.g. cells with no treatment applied when testing cellular biological perturbations. Using the unconditional marginal as the reference would generally be misspecified for this problem.

In the Bayes setting, one begins with a common joint distribution and conditionals $P(x \mid c_1)$ and $P(x \mid c_2)$. Under the relevant conditional-independence assumption,

$$P(x \mid c_1, c_2) \propto \frac{P(x \mid c_1)P(x \mid c_2)}{P(x)},$$

so the unconditional marginal $P(x)$ is indeed the appropriate denominator. By contrast, in many settings of interest, a_1 and a_2 represent interventions rather than conditional events arising from a common joint distribution, and P^0 is a control distribution that need not be restricted to be the unconditional marginal or a mixture such as $\sum_a P(a)P^a$.

Indeed, under the modular intervention assumptions in our causal motivation (Sec. 2 and App. B), the control distribution is instead the appropriate reference. For example, suppose $x = (x_1, x_2)$ and

$$P^0(x) = P^0(x_1)P^0(x_2 \mid x_1).$$

If a_1 changes only the first mechanism and a_2 changes only the second, then

$$P^{a_1}(x) = P^{a_1}(x_1)P^0(x_2 \mid x_1), \qquad P^{a_2}(x) = P^0(x_1)P^{a_2}(x_2 \mid x_1).$$

Modularity gives the combined intervention

$$P^{a_1,a_2}(x) = P^{a_1}(x_1)P^{a_2}(x_2 \mid x_1) = \frac{P^{a_1}(x)P^{a_2}(x)}{P^0(x)}.$$

Intuitively, the control distribution appears because it contains the mechanisms that are shared by the single interventions and must be cancelled when their effects are combined.

Bradley et al. (2025) also give a complementary mathematical argument from the perspective of projective composition. Consider image generation with two distributions P^{a_1} and P^{a_2}, each containing one object in a distinct region, and an unconditional training distribution P^u that also contains only one object. A Bayes composition

$$P(x) \propto \frac{P^{a_1}(x)P^{a_2}(x)}{P^u(x)}$$

cannot generate the desired two-object samples in an idealized support setting: $P(x) = 0$ whenever either $P^{a_1}(x) = 0$ or $P^{a_2}(x) = 0$, so its support remains contained in the intersection of the single-object supports. Equivalently, if P^{a_1} and P^{a_2} are interpreted as conditionals $P(x \mid c_1)$ and $P(x \mid c_2)$ of the same underlying distribution, then the desired conditional $P(x \mid c_1, c_2)$ is not well-defined in this setting, since neither conditional has support on images with two objects and the joint event (c_1, c_2) has zero probability.

An empty-background reference permits a different construction. If each P^{a_i} differs from a shared background P^0 only on a disjoint coordinate block M_i, then

$$P^{a_i}(x) = P^{a_i}(x_{M_i})P^0(x_{M_i^c}),$$

and factorization of the background causes the composition

$$P^0(x) \prod_i \frac{P^{a_i}(x)}{P^0(x)}$$

to simplify to a distribution that retains the background outside the object regions and independently inserts the marginal for each M_i. The reference distribution encodes the shared background relative to which the interventions factorize, whereas an arbitrary unconditional marginal need not have this property.

G Proofs and additional theoretical details

For clarity, throughout this section we specialize to a base composition with two perturbation sources, $\mathcal{A}_o = \{0, a_1, a_2\}$, and write

$$P^w(x) = \frac{1}{Z^w} \frac{P^{a_1}(x) P^{a_2}(x)}{P^0(x)}, \qquad Z^w = \int \frac{P^{a_1}(x) P^{a_2}(x)}{P^0(x)}\, dx.$$

Additionally, throughout this section, P^w denotes the exact composition of the true source distributions and $\widehat{P}^w$ the exact composition of perturbed (i.e., learned) source distributions.

G.1 Preliminaries: Gaussian weighted composition

As a preliminary, we prove a necessary fact about weighted compositions of Gaussians.

Proposition 2. *Let $P^a = \mathcal{N}(\mu_a, \Sigma_a)$ for $a \in \mathcal{A}_o$. Then, w is a valid weighting function if and only if $\sum_{a \in \mathcal{A}_o} w_a \Sigma_a^{-1} \succ 0$. In this case, $\mathbf{P}^w(x) = \mathcal{N}(\mu^w, \Sigma^w)$ for*

$$\Sigma^w = \left(\sum\nolimits_{a \in \mathcal{A}_o} w_a \Sigma_a^{-1} \right)^{-1}, \qquad \mu^w = \Sigma^w \sum\nolimits_{a \in \mathcal{A}_o} w_a \Sigma_a^{-1} \mu_a. \tag{11}$$

Proof. Let $x \in \mathbb{R}^d$. The probability density function for each multivariate Gaussian P^a is given by:

$$P^a(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_a|}} \exp\left(-\frac{1}{2} (x - \mu_a)^\top \Sigma_a^{-1} (x - \mu_a) \right)$$

We wish to evaluate $f(x)$, the product of these densities raised to their respective weights w_a:

$$\begin{aligned} f(x) &= \prod_{a \in \mathcal{A}_o} P^a(x)^{w_a} \\ &= \prod_{a \in \mathcal{A}_o} \left((2\pi)^{-d/2} |\Sigma_a|^{-1/2} \exp\left(-\frac{1}{2} (x - \mu_a)^\top \Sigma_a^{-1} (x - \mu_a) \right) \right)^{w_a} \\ &= \left(\prod_{a \in \mathcal{A}_o} (2\pi)^{-d w_a / 2} |\Sigma_a|^{-w_a/2} \right) \exp\left(-\frac{1}{2} \sum_{a \in \mathcal{A}_o} w_a (x - \mu_a)^\top \Sigma_a^{-1} (x - \mu_a) \right). \end{aligned}$$

Let $M = \sum_{a \in \mathcal{A}_o} w_a \Sigma_a^{-1}$, and let $b = \sum_{a \in \mathcal{A}_o} w_a \Sigma_a^{-1} \mu_a$. Absorbing terms independent of x into a constant C, we have

$$f(x) = C \exp\left(-\frac{1}{2} \left(x^\top M x - 2 x^\top b \right) \right).$$

Assuming $M \succ 0$, completing the square gives

$$f(x) \propto \exp\left(-\frac{1}{2} (x - M^{-1} b)^\top M (x - M^{-1} b) \right).$$

We can recognize this as the kernel of a multivariate Gaussian distribution with mean $M^{-1} b$ and precision M, which is well-known to be integrable if and only if $M \succ 0$ (that is, M is positive definite). Therefore, assuming $M \succ 0$, normalizing $f(x)$ yields $P^w(x) = \mathcal{N}(\mu^w, \Sigma^w)$ for

$$\Sigma^w = \left(\sum_{a \in \mathcal{A}_o} w_a \Sigma_a^{-1} \right)^{-1}, \qquad \mu^w = \Sigma^w \sum_{a \in \mathcal{A}_o} w_a \Sigma_a^{-1} \mu_a.$$

□

G.2 Arbitrary instability under vanishing source error

Proposition 1 states the composition operator is not uniformly stable under small errors in its source distributions. This can occur even for one-dimensional Gaussians. We first give a simple explicit construction which immediately proves Proposition 1, and then show that the same distributions are also more generically unstable to independent random perturbations.

Proposition 3 (Vanishing source error with catastrophic composition error)**.** *For $n \geq 2$, let the true source distributions be centered one-dimensional Gaussians with precisions*

$$\lambda^0 = 2, \qquad \lambda^{a_1} = \lambda^{a_2} = 1 + \frac{1}{2n},$$

and let the estimated sources differ only in the base precision,

$$\widehat{\lambda}^0 = 2 + \frac{1}{n} - \frac{1}{n^2}, \qquad \widehat{\lambda}^{a_1} = \lambda^{a_1}, \qquad \widehat{\lambda}^{a_2} = \lambda^{a_2}.$$

Then the source-distribution errors vanish as $n \to \infty$, including in TV, KL, W_2, and relative L^2. Simultaneously, the corresponding base compositions satisfy

$$\mathrm{TV}(P_n^w, \widehat{P}_n^w) \to 1, \qquad D_{\mathrm{KL}}(P_n^w \| \widehat{P}_n^w) \to \infty, \qquad W_2(P_n^w, \widehat{P}_n^w) \to \infty.$$

Proof. For centered Gaussians, base composition adds numerator precisions and subtracts the base precision. That is,

$$\lambda_n^w = \lambda^{a_1} + \lambda^{a_2} - \lambda^0 = \frac{1}{n}, \qquad \widehat{\lambda}_n^w = \widehat{\lambda}^{a_1} + \widehat{\lambda}^{a_2} - \widehat{\lambda}^0 = \frac{1}{n^2}.$$

Thus

$$P_n^w = \mathcal{N}(0, n), \qquad \widehat{P}_n^w = \mathcal{N}(0, n^2).$$

The only perturbed source precision differs from its true value by $O(n^{-1})$, while both precisions remain uniformly bounded away from zero and infinity. Consequently, the corresponding Gaussian source distributions converge in TV, KL, W_2, and χ^2, so in particular

$$\left\| \frac{\widehat{P}_n^0 - P_n^0}{P_n^0} \right\|_{L^2(P_n^0)}^2 = \chi^2(\widehat{P}_n^0 \| P_n^0) \to 0.$$

Their score difference is

$$\nabla \log \widehat{P}_n^0(x) - \nabla \log P_n^0(x) = -\left(\frac{1}{n} - \frac{1}{n^2}\right) x,$$

and therefore

$$\|\nabla \log \widehat{P}_n^0 - \nabla \log P_n^0\|_{L^2(P_n^0)} = O(n^{-1}).$$

For the compositions,

$$W_2(P_n^w, \widehat{P}_n^w) = \left|\sqrt{n} - n\right| \to \infty,$$

and

$$D_{\mathrm{KL}}(P_n^w \| \widehat{P}_n^w) = \frac{1}{2}\left(\frac{1}{n} - 1 + \log n\right) \to \infty.$$

Finally, scale invariance of TV gives

$$\mathrm{TV}(P_n^w, \widehat{P}_n^w) = \mathrm{TV}(\mathcal{N}(0,1), \mathcal{N}(0,n)) \to 1.$$

□

One may worry that the above example is adversarial, and that such ill-conditioning of would not hold under mild changes to the setup. However, catastrophic instability does not require a specially aligned error direction, as the following example demonstrates.

Proposition 4 (Independent random perturbations also yield catastrophic composition error)**.** *Consider the same true source distributions as in Proposition 3. Draw*

$$\xi_0, \xi_{a_1}, \xi_{a_2} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$$

once, and let $\epsilon_n = n^{-1/4}$*. Define estimated source distributions by leaving all source precisions unchanged and independently perturbing their means:*

$$\widehat{P}_n^i = \mathcal{N}\big(\epsilon_n \xi_i, \lambda_i^{-1}\big), \qquad i \in \{0, a_1, a_2\}.$$

Then, with probability one, every individual source error converges to zero in TV, KL, W_2*, and relative* L^2*, while*

$$\mathrm{TV}(P_n^w, \widehat{P}_n^w) \to 1, \qquad W_2(P_n^w, \widehat{P}_n^w) \to \infty.$$

Proof. Because only the means are perturbed, for every $i \in \{0, a_1, a_2\}$,

$$W_2(P_n^i, \widehat{P}_n^i) = \epsilon_n |\xi_i| \to 0.$$

For Gaussians with equal variance,

$$D_{\mathrm{KL}}(\widehat{P}_n^i \| P_n^i) = \frac{\lambda_i \epsilon_n^2 \xi_i^2}{2} = O(n^{-1/2}),$$

so $\mathrm{TV}(P_n^i, \widehat{P}_n^i) \to 0$ as well by Pinsker's inequality, and

$$\left\| \frac{\widehat{P}_n^i - P_n^i}{P_n^i} \right\|_{L^2(P_n^i)}^2 = \chi^2(\widehat{P}_n^i \| P_n^i) = \exp\big(\lambda_i \epsilon_n^2 \xi_i^2\big) - 1 \to 0.$$

The source precisions are unchanged, so both the true and perturbed compositions have precision

$$\lambda_n^w = \widehat{\lambda}_n^w = \lambda^{a_1} + \lambda^{a_2} - \lambda^0 = \frac{1}{n}.$$

Thus both compositions have variance n. For a one-dimensional Gaussian, the natural parameter associated with the mean is $\lambda\mu$. Therefore the perturbed composed mean is

$$\widehat{\mu}_n^w = \frac{\lambda^{a_1} \epsilon_n \xi_{a_1} + \lambda^{a_2} \epsilon_n \xi_{a_2} - \lambda^0 \epsilon_n \xi_0}{\lambda^{a_1} + \lambda^{a_2} - \lambda^0}.$$

Substituting the source precisions gives

$$\widehat{\mu}_n^w = n^{3/4} \left[\left(1 + \frac{1}{2n}\right)(\xi_{a_1} + \xi_{a_2}) - 2\xi_0 \right].$$

Let $C = \xi_{a_1} + \xi_{a_2} - 2\xi_0$. Since C is a nondegenerate Gaussian random variable, $P(C = 0) = 0$. Hence, with probability one,

$$\left(1 + \frac{1}{2n}\right)(\xi_{a_1} + \xi_{a_2}) - 2\xi_0 \longrightarrow C \neq 0,$$

and therefore

$$|\widehat{\mu}_n^w| \sim |C| n^{3/4} \to \infty.$$

Because the two compositions have the same variance,

$$W_2(P_n^w, \widehat{P}_n^w) = |\widehat{\mu}_n^w| \to \infty.$$

Moreover, their common standard deviation is $\sqrt{n}$, while their standardized mean separation satisfies

$$\frac{|\widehat{\mu}_n^w|}{\sqrt{n}} = n^{1/4} \left| \left(1 + \frac{1}{2n}\right)(\xi_{a_1} + \xi_{a_2}) - 2\xi_0 \right| \to \infty$$

almost surely. For equal-variance one-dimensional Gaussians,

$$\mathrm{TV}\big(\mathcal{N}(0, \sigma^2), \mathcal{N}(\mu, \sigma^2)\big) = 2\Phi\left(\frac{|\mu|}{2\sigma}\right) - 1,$$

where Φ is the standard normal CDF. It follows that

$$\mathrm{TV}(P_n^w, \widehat{P}_n^w) \to 1$$

almost surely. □

G.3 Local sensitivity to relative L^2 source error

This section contains the proof of Theorem 1 and a tightness result. For ease of reading, we reproduce the definitions from Sec. 3 here. Define the relative source error

$$u_i(x) = \frac{\widehat{P}^i(x) - P^i(x)}{P^i(x)}, \qquad i \in \{0, a_1, a_2\}.$$

Theorem 1 (First-order sensitivity bound)**.** *For $s \in [0,1]$, consider the convex interpolation $P_s^i = P^i(1+su_i) = (1-s)P^i + s\widehat{P}^i$ and let P_s^w denote the corresponding base composition whenever it is valid. Define source-to-composition ratios $R_i(x) = \mathbf{P}^w(x)P^i(x)^{-1}$ and*

$$\kappa_i := \left(\mathbb{E}_{P^i}[R_i(x)^2]\right)^{1/2} = \sqrt{1+\chi^2(\mathbf{P}^w\|P^i)}.$$

Assuming P_s^w is well-defined for all sufficiently small $s \geq 0$ and $s \mapsto P_s^w$ is right-differentiable in L^1 at $s=0$, then as $s \downarrow 0$,

$$\mathrm{TV}(\mathbf{P}^w, P_s^w) \leq s \sum\nolimits_{i\in\{0,a_1,a_2\}} \kappa_i \|u_i\|_{L^2(P^i)} + o(s).$$

Proof. Let

$$f_s(x) = \frac{P_s^{a_1}(x)P_s^{a_2}(x)}{P_s^0(x)}, \qquad Z_s^w = \int f_s(x)\,dx,$$

so that $P_s^w = f_s/Z_s^w$. Since $P_s^i = P^i(1+su_i)$,

$$\partial_s \log P_s^i(x)\big|_{s=0} = u_i(x).$$

Therefore

$$\partial_s \log f_s(x)|_{s=0} = u_{a_1}(x) + u_{a_2}(x) - u_0(x) =: H(x),$$

and hence

$$\partial_s f_s(x)|_{s=0} = f_0(x)H(x).$$

Differentiating the normalizing constant gives

$$\frac{\partial_s Z_s^w|_{s=0}}{Z_0^w} = \frac{1}{Z_0^w}\int f_0(x)H(x)\,dx = \mathbb{E}_{P^w}[H].$$

Finally, differentiating $P_s^w = f_s/Z_s^w$ yields

$$\partial_s P_s^w(x)|_{s=0} = P^w(x)\left(H(x) - \mathbb{E}_{P^w}[H]\right).$$

Hence

$$\mathrm{TV}(\mathbf{P}^w, P_s^w) = \frac{s}{2}\mathbb{E}_{\mathbf{P}^w}\left[|H - \mathbb{E}_{\mathbf{P}^w}H|\right] + o(s) \leq s\sum_i \mathbb{E}_{\mathbf{P}^w}|u_i| + o(s).$$

For each source, change of measure followed by Cauchy-Schwarz gives

$$\mathbb{E}_{\mathbf{P}^w}|u_i| = \mathbb{E}_{P^i}[R_i|u_i|] \leq \left(\mathbb{E}_{P^i}[R_i^2]\right)^{1/2}\|u_i\|_{L^2(P^i)} = \kappa_i\|u_i\|_{L^2(P^i)}.$$

Summing the three terms proves the claim. □

Note that the same argument extends directly to arbitrary weighted compositions $\mathbf{P}^w \propto \prod_i (P^i)^{w_i}$, in which case the first-order perturbation becomes $\sum_i w_i u_i$ and the bound is

$$\mathrm{TV}(\mathbf{P}^w, P_s^w) \leq s\sum_i |w_i|\kappa_i\|u_i\|_{L^2(P^i)} + o(s).$$

However, small endpoint relative error and finite κ_i do not, by themselves, imply a global finite-perturbation bound, since local denominator errors can still create arbitrarily large reciprocal errors. A stronger global statement is given in App. G.9.

Note that Theorem 1 implicitly assumes $u_i \in L^2(P^i)$. That is, we assume the learned model has finite relative L^2 error: $\int \frac{(\hat{p}_i - p_i)^2}{p_i} < \infty$. Otherwise, the bound is vacuous.

The following example shows that the dependence on κ_i in Theorem 1 is tight up to constant factors.

Proposition 5 (Tightness of the local sensitivity bound)**.** *The dependence on κ_i in Theorem 1 is tight up to a constant factor. In particular, there exist source distributions P^0, P^{a_1}, P^{a_2} such that*

$$\lim_{s\downarrow 0} \frac{\mathrm{TV}(P^w, P^w_s)}{s} = \frac{1}{2}\kappa_{a_1},$$

while κ_{a_1} can be arbitrarily large.

Proof. Let $\mathcal{X} = [0,1]$, fix $\delta \in (0, 1/2)$, and set $A = [0, \delta]$. Take $P^0 = P^{a_1} = \mathrm{Unif}[0,1]$ and

$$P^{a_2}(x) = \frac{\mathbb{1}_A(x)}{2\delta} + \frac{\mathbb{1}_{A^c}(x)}{2(1-\delta)}.$$

Then $P^w = P^{a_2}$ and $P^w(A) = P^w(A^c) = 1/2$. Perturb only P^{a_1} with

$$u_{a_1}(x) = \frac{\mathbb{1}_A(x) - \delta}{\sqrt{\delta(1-\delta)}}.$$

Under P^{a_1}, $\mathbb{E}[u_{a_1}] = 0$ and $\|u_{a_1}\|_{L^2(P^{a_1})} = 1$. We have that $u_{a_1} > -1$ for $\delta < 1/2$, so $\widehat{P}^{a_1} = P^{a_1}(1 + u_{a_1})$ is a valid density. Also,

$$\kappa_{a_1} = \left(\mathbb{E}_{P^{a_1}}\left[\left(\frac{P^w}{P^{a_1}}\right)^2\right]\right)^{1/2} = \frac{1}{2\sqrt{\delta(1-\delta)}}.$$

For convenience of notation, let

$$a = \sqrt{\frac{1-\delta}{\delta}}, \qquad b = \sqrt{\frac{\delta}{1-\delta}},$$

we have $u_{a_1} = a$ on A and $u_{a_1} = -b$ on A^c. Then

$$P^w_s(A) = \frac{\frac{1}{2}(1 + sa)}{1 + s(a-b)/2},$$

and therefore

$$\mathrm{TV}(P^w, P^w_s) = \left|P^w_s(A) - \frac{1}{2}\right| = \frac{s(a+b)}{4\,(1 + s(a-b)/2)}.$$

Since $a + b = 1/\sqrt{\delta(1-\delta)}$,

$$\lim_{s\downarrow 0} \frac{\mathrm{TV}(P^w, P^w_s)}{s} = \frac{1}{4\sqrt{\delta(1-\delta)}} = \frac{1}{2}\kappa_{a_1}.$$

Finally, $\kappa_{a_1} \to \infty$ as $\delta \downarrow 0$, proving the claim. □

G.4 Inference-time approximation error in Gaussians

In this section, we prove that even for simple and analytically tractable Gaussians, the composition operator does not commute with noising, which leads the naïve denoising process to sample from the incorrect distribution. We further provide corollary observations on the error-sensitivity of naive composition and exact composition in the presence of estimation error.

Proposition 6. *Let $P^a = \mathcal{N}(0, \Sigma_a)$ for $a \in \mathcal{A}_o$, and assume $\{\Sigma_a\}_{a\in\mathcal{A}_o}$ are mutually commutative and that $\sum_{a\in\mathcal{A}_o} w_a = 1$, and assume the noising schedule satisfies the idealized terminal conditions $\alpha_1 = 0$ and $\gamma_1 = 1$, as in the idealized limit of standard variance-preserving diffusion. Then, the naïve denoising ODE (8) yields $\widetilde{P}^{w,\mathrm{nvo}} = \mathcal{N}(0, \Sigma^{w,\mathrm{nvo}})$, where*

$$\Sigma^{w,\mathrm{nvo}} = \prod\nolimits_{a\in\mathcal{A}_o} \Sigma_a^{w_a}.$$

To prove this, we first begin by proving a necessary Lemma.

Lemma 1 (Log-derivative of a noised Gaussian covariance)**.** *For $P^a = \mathcal{N}(0, \Sigma_a)$, the noising marginal is $P_t^a = \mathcal{N}(0, \Sigma_t^a)$ with $\Sigma_t^a = \alpha_t^2 \Sigma_a + \gamma_t^2 I$, and*

$$\frac{d}{dt} \ln \Sigma_t^a = 2\mu_t I + \sigma_t^2 (\Sigma_t^a)^{-1}.$$

Proof. The first claim follows directly from the noising kernel $\text{Noise}_t(x_0) = \mathcal{N}(\alpha_t x_0; \gamma_t^2 I)$. For the second, both Σ_t^a and

$$\dot{\Sigma}_t^a = 2\alpha_t \dot{\alpha}_t \Sigma_a + 2\gamma_t \dot{\gamma}_t I$$

are linear combinations of Σ_a and I, so they commute, giving $\frac{d}{dt} \ln \Sigma_t^a = \dot{\Sigma}_t^a (\Sigma_t^a)^{-1}$. Using $\mu_t = \dot{\alpha}_t / \alpha_t$ and $\sigma_t^2 = 2\gamma_t \dot{\gamma}_t - 2\mu_t \gamma_t^2$,

$$\dot{\Sigma}_t^a = 2\mu_t(\alpha_t^2 \Sigma_a + \gamma_t^2 I) + \sigma_t^2 I = 2\mu_t \Sigma_t^a + \sigma_t^2 I.$$

Right-multiplying by $(\Sigma_t^a)^{-1}$ gives the result. □

Now we begin the proof of Proposition 6.

Proof. Since each P^a is mean-zero Gaussian, $S_t^a(x) = -(\Sigma_t^a)^{-1} x$, so the naïvely composed score is linear in x:

$$S_t^{w,\text{nv}}(x) = \sum_{a \in \mathcal{A}_\text{o}} w_a S_t^a(x) = -J_t^{w,\text{nv}}\, x, \qquad J_t^{w,\text{nv}} := \sum_{a \in \mathcal{A}_\text{o}} w_a (\Sigma_t^a)^{-1}.$$

Combined with $u_t(x) = \mu_t x$, the naïve denoising velocity field is also linear in x:

$$\tilde{v}_t^{w,\text{nv}}(x) = A_t\, x, \qquad A_t := \mu_t I + \tfrac{1}{2}\sigma_t^2 J_t^{w,\text{nv}},$$

with A_t symmetric. The naïve denoising ODE (8), integrated from $t = 1$ to $t = 0$, is therefore

$$\frac{dx_t}{dt} = A_t x_t, \qquad x_1 \sim \mathcal{N}(0, I),$$

where the initial distribution reflects $\mathbf{P}_1^w = \mathcal{N}(0, I)$.

Commutativity. Note that every matrix appearing in this proof, Σ_t^a, $(\Sigma_t^a)^{-1}$, A_t, Σ_t, is built from $\{\Sigma_a\}_{a \in \mathcal{A}_\text{o}}$ and I by sums, products, inverses, and (for Σ_t) matrix exponentials. Since the Σ_a pairwise commute by hypothesis, any two such matrices commute. In particular, Σ_t commutes with A_t and with $\dot{\Sigma}_t$, which justifies all matrix manipulations below and the usage of the identity $\frac{d}{dt} \ln \Sigma_t = \dot{\Sigma}_t \Sigma_t^{-1}$ previously.

Covariance evolution. The covariance satisfies $\dot{\Sigma}_t = A_t \Sigma_t + \Sigma_t A_t^\top = 2A_t \Sigma_t$ (Särkkä & Solin, 2019, Equation (6.2)), so

$$\frac{d}{dt} \ln \Sigma_t = 2A_t = 2\mu_t I + \sigma_t^2 \sum_{a \in \mathcal{A}_\text{o}} w_a (\Sigma_t^a)^{-1}.$$

Using $\sum_{a \in \mathcal{A}_\text{o}} w_a = 1$ to absorb the $2\mu_t I$ term into the sum, then applying Lemma 1 termwise,

$$\frac{d}{dt} \ln \Sigma_t = \sum_{a \in \mathcal{A}_\text{o}} w_a \big[2\mu_t I + \sigma_t^2 (\Sigma_t^a)^{-1}\big] = \sum_{a \in \mathcal{A}_\text{o}} w_a \frac{d}{dt} \ln \Sigma_t^a.$$

Integration and boundary conditions. Integrating from $t = 1$ to $t = 0$,

$$\ln \Sigma_0 - \ln \Sigma_1 = \sum_{a \in \mathcal{A}_\text{o}} w_a \big(\ln \Sigma_0^a - \ln \Sigma_1^a\big).$$

The boundary values $\alpha_0 = 1, \gamma_0 = 0$ give $\Sigma_0^a = \Sigma_a$, and $\alpha_1 = 0, \gamma_1 = 1$ give $\Sigma_1^a = \Sigma_1 = I$, so $\ln \Sigma_1^a = \ln \Sigma_1 = 0$. Hence

$$\ln \Sigma_0 = \sum_{a \in \mathcal{A}_\text{o}} w_a \ln \Sigma_a, \qquad \Sigma_0 = \exp\Big(\sum_{a \in \mathcal{A}_\text{o}} w_a \ln \Sigma_a\Big) = \prod_{a \in \mathcal{A}_\text{o}} \Sigma_a^{w_a},$$

where the last equality uses that for commuting symmetric positive-definite matrices, $\exp(\ln A + \ln B) = AB$.

Mean. The mean m_t satisfies $\dot{m}_t = A_t m_t$ with $m_1 = 0$, so $m_t \equiv 0$. Thus, $\widetilde{P}^{w,\mathrm{nvo}} = \mathcal{N}\big(0, \prod_{a\in\mathcal{A}_o} \Sigma_a^{w_a}\big)$. □

In this case, as a result of this approximation error, addition and subtraction are effectively replaced by multiplication and division after the denoising process. These resulting quantities can in fact be *arbitrarily far from each other*, as a simple example demonstrates: Let $|\mathcal{A}_o| = 2$, $w_1 = w_2 = \frac{1}{2}$, and $\Sigma_1 = \varepsilon I$, $\Sigma_2 = \varepsilon^{-1} I$ for $\varepsilon > 0$. Then

$$\Sigma^{w,\mathrm{nvo}} = I \qquad \text{and} \qquad \Sigma^{w} = \frac{2\varepsilon}{1+\varepsilon^2} I,$$

so

$$\|\Sigma^{w,\mathrm{nvo}}\|_{\mathrm{op}} / \|\Sigma^{w}\|_{\mathrm{op}} = (1+\varepsilon^2)/(2\varepsilon) \to \infty \quad \text{as} \quad \varepsilon \to 0^+$$

Even though this demonstrates that naïve denoising composition can be catastrophically unstable, it by the same token demonstrates that, on estimated distributions, naïve denoising composition can be more robust to estimation error than exact composition, provided it is not unstable on the true distributions.

Corollary 1.1 (Naïve composition can be less sensitive to estimation error)**.** *Consider the Factorized Conditional Gaussian sources*

$$P^0 = \mathcal{N}(0, I_2), \qquad P^{a_1} = \mathcal{N}(0, \mathrm{diag}(n, 1)), \qquad P^{a_2} = \mathcal{N}(0, \mathrm{diag}(1, n)),$$

whose true base composition is $\mathbf{P}^w = \mathcal{N}(0, nI_2)$. Suppose only the base distribution is perturbed to $\hat{P}^0 = \mathcal{N}(0, \mathrm{diag}(1+\varepsilon_1, 1+\varepsilon_2))$, and $\hat{P}^{a_1} = P^{a_1}$, $\hat{P}^{a_2} = P^{a_2}$. Then the exact base composition of the perturbed sources is

$$\hat{\mathbf{P}}^w = \mathcal{N}(0, \mathrm{diag}(\tau_1, \tau_2)), \qquad \tau_j = \frac{n(1+\varepsilon_j)}{1+(n+1)\varepsilon_j}.$$

With the same assumptions as Proposition 6, the naïve denoising ODE yields

$$\widetilde{P}^w = \mathcal{N}(0, \mathrm{diag}(\nu_1, \nu_2)), \qquad \nu_j = \frac{n}{1+\varepsilon_j}.$$

Hence, for fixed $\varepsilon_1, \varepsilon_2 > 0$ and $n \to \infty$,

$$\tau_j = \Theta(1), \qquad \nu_j = \Theta(n).$$

Thus, true composition has variance $\Theta(n)$ in each coordinate, and the variance in the exact composition of the estimated source distributions collapses to $\Theta(1)$, while naïve composition preserves the correct variance scale. Consequently,

$$\mathrm{TV}(\mathbf{P}^w, \hat{\mathbf{P}}^w) \to 1, \qquad \mathrm{TV}(\mathbf{P}^w, \widetilde{P}^w) < 1,$$

and naïve composition is strictly closer in TV to the true composition than the exact composition of the estimated distributions, for sufficiently large n.

Corollary 1.2 (Naïve composition can be more stable under vanishing source error)**.** *For $n \geq 2$, consider*

$$P_n^0 = \mathcal{N}(0, I_2), \qquad P_n^{a_1} = \mathcal{N}(0, \mathrm{diag}(n, 1+n^{-2})), \qquad P_n^{a_2} = \mathcal{N}(0, \mathrm{diag}(1+n^{-2}, n)).$$

These distributions are not Factorized Conditionals for any finite n, but approach Factorized Conditionals as $n \to \infty$. By (11), their base composition is

$$\mathbf{P}_n^w = \mathcal{N}(0, s_n I_2), \qquad s_n = \frac{n(1+n^{-2})}{1-(n-1)n^{-2}} = \Theta(n).$$

Perturb only the base distribution to $\widehat{P}_n^0 = \mathcal{N}(0, (1+n^{-1/2})I_2)$, and let $\widehat{P}_n^{a_1} = P_n^{a_1}$, $\widehat{P}_n^{a_2} = P_n^{a_2}$. The source-wise TV errors vanish as $n \to \infty$. The exact base composition of the estimated sources is

$$\widehat{\mathbf{P}}_n^w = \mathcal{N}(0, \widehat{s}_n I_2), \qquad \widehat{s}_n = \frac{n(1+n^{-2})(1+n^{-1/2})}{1-(n-1)n^{-2}+(n+1)n^{-1/2}+n^{-5/2}} = \Theta(\sqrt{n}).$$

In contrast, by Proposition 6 the naïve composition is

$$\widetilde{P}_n^w = \mathcal{N}(0, \widetilde{s}_n I_2), \qquad \widetilde{s}_n = \frac{n(1+n^{-2})}{1+n^{-1/2}},$$

for which $\widetilde{s}_n / s_n \to 1$. Consequently,

$$\mathrm{TV}(\mathbf{P}_n^w, \widehat{\mathbf{P}}_n^w) \to 1, \qquad \mathrm{TV}(\mathbf{P}_n^w, \widetilde{P}_n^w) \to 0.$$

It is important to point out a contrast. In this fixed-dimensional diagonal Gaussian setting, naïve ODE composition is stable to vanishing source-wise TV perturbations. That is, if $\mathrm{TV}(P_n^a, \widehat{P}_n^a) \to 0$ for every source distribution, then the naïve compositions obtained from the true distributions and estimated distributions also converge in TV. This follows because vanishing source-wise TV error forces each variance ratio between corresponding estimated and true distributions to converge to 1, while the naïve ODE covariance is a product of the source distribution covariances, so its corresponding covariance ratios also converge to 1. Thus, vanishing estimation error cannot itself cause naïve composition to become catastrophically inaccurate, and any such arbitrarily large asymptotic failure must already occur without estimation error. Meanwhile, exact composition on the estimated distributions need not satisfy this form of stability, as Corollary 1.2 demonstrates.

G.5 Feynman-Kac correctors under factorized conditionals

In this section, we prove the claim from the text that the Feynman-Kac Corrector log-weight drift $g_t(x)$ vanishes under Factorized Conditionals and that the FKC SMC instantiation coincides with the naïve denoising ODE.

Proposition 7 (Feynman-Kac Correctors under factorized conditionals)**.** *Assume the set of distributions $(P^0, P^{a_1}, \dots P^{a_k})$ are Factorized Conditional. Then, for the base-composition on $(P^0, P^{a_1}, \dots P^{a_k})$, the Feynman-Kac Corrector log-weight drift vanishes:*

$$g_t(x) = 0 \;\; \textit{for all} \;\; x \in \mathbb{R}^n, \; t \in [0,1],$$

and the Feynman-Kac Corrector denoising process coincides exactly with the naïve denoising SDE.

To begin, we first prove a simple lemma that allows us to not worry about the time t.

Lemma 2 (Coordinate-wise noising preserves factorization)**.** *If P_0 on $\mathbb{R}^n$ factorizes over a partition $[n] = \bigsqcup_\ell B_\ell$ as $P_0(x) = \prod_\ell P_0(x_{B_\ell})$, and the noising SDE acts coordinate-wise, then for every $t \in [0,1]$ the noising marginal P_t factorizes over the same partition: $P_t(x) = \prod_\ell P_t(x_{B_\ell})$.*

Proof. The coordinate-wise noising kernel satisfies $(\mathrm{Noise}_t(x_0))_{B_\ell} = \alpha_t (x_0)_{B_\ell} + \gamma_t Z_{B_\ell}$ with $Z_{B_\ell} \sim \mathcal{N}(0, I_{|B_\ell|})$ independent across ℓ. If $X_0 \sim P_0$ has independent blocks by hypothesis, then the noised blocks $(X_t)_{B_\ell}$ are functions of disjoint independent inputs and remain mutually independent, so the joint density of X_t factorizes block-wise. □

Now, we prove Proposition 7.

Proof. For convenience, let $\mathcal{A}_{\mathrm{o}}^- := \mathcal{A}_{\mathrm{o}} \setminus \{0\}$, and let the distributions $\{P^0\} \cup \{P^a\}_{a \in \mathcal{A}_{\mathrm{o}}}$ on $\mathbb{R}^n$ denote the Factorized Conditional distributions.

Let $k := |\mathcal{A}_{\mathrm{o}}^-|$. The weights satisfy $w_0 + \sum_{a \in \mathcal{A}_{\mathrm{o}}^-} w_a = (1-k) + k = 1$, so the divergence term in (10) vanishes. It remains to show

$$\Big\| w_0 S_t^0 + \sum_{a \in \mathcal{A}_{\mathrm{o}}^-} w_a S_t^a \Big\|^2 \; - \; \Big(w_0 \|S_t^0\|^2 + \sum_{a \in \mathcal{A}_{\mathrm{o}}^-} w_a \|S_t^a\|^2 \Big) \; = \; 0. \tag{12}$$

Block decomposition of scores. The noising operator Noise_t acts coordinate-wise because μ_t and σ_t are scalar, so by Lemma 2, the factorization (4) is preserved at every t. We drop the t subscript hereafter.

Define the following vectors in $\mathbb{R}^n$, each supported on a single block of the partition:

$$\begin{aligned}
\tilde{u}_0 &: \text{ supported on } M_0, \quad (\tilde{u}_0)_{M_0} = \nabla_{x_{M_0}} \log P^0(x_{M_0}); \\
\tilde{u}_a &: \text{ supported on } M_a, \quad (\tilde{u}_a)_{M_a} = \nabla_{x_{M_a}} \log P^0(x_{M_a}), \quad a \in \mathcal{A}_{\mathrm{o}}^-; \\
\tilde{v}_a &: \text{ supported on } M_a, \quad (\tilde{v}_a)_{M_a} = \nabla_{x_{M_a}} \log P^a(x_{M_a}), \quad a \in \mathcal{A}_{\mathrm{o}}^-.
\end{aligned}$$

Differentiating the $P^0(x) = P^0(x|_{M_0}) \prod_{a \in \mathcal{A}_{\mathrm{o}}^-} P^0(x|_{M_a})$ term from (4) block-wise yields

$$S^0 \; = \; \tilde{u}_0 + \sum_{a \in \mathcal{A}_{\mathrm{o}}^-} \tilde{u}_a. \tag{13}$$

For each $a \in \mathcal{A}_o^-$, expanding $\log P^a(x) = \log P^a(x_{M_a}) + \log P^0(x_{M_0}) + \sum_{a' \neq a} \log P^0(x_{M_{a'}})$ via (4) and differentiating gives

$$S^a = \tilde{v}_a + \tilde{u}_0 + \sum_{a' \in \mathcal{A}_o^- \setminus \{a\}} \tilde{u}_{a'}. \tag{14}$$

The blocks $\{M_0\} \cup \{M_a\}_{a \in \mathcal{A}_o^-}$ are all mutually disjoint, so vectors supported on different blocks are orthogonal in $\mathbb{R}^n$. The vectors $\tilde{u}_a$ and $\tilde{v}_a$ share support M_a and are not in general orthogonal, but $\tilde{u}_a$ never appears in S^a, so this non-orthogonality plays no role in the calculations below.

Composed score. Substituting (13) and (14) and collecting coefficients block-by-block,

$$\begin{aligned} w_0 S^0 + \sum_{a \in \mathcal{A}_o^-} w_a S^a &= (1-k)\left(\tilde{u}_0 + \sum_{a \in \mathcal{A}_o^-} \tilde{u}_a\right) + \sum_{a \in \mathcal{A}_o^-} \left(\tilde{v}_a + \tilde{u}_0 + \sum_{a' \neq a} \tilde{u}_{a'}\right) \\ &= \big[(1-k) + k\big]\, \tilde{u}_0 + \sum_{a \in \mathcal{A}_o^-} \big[(1-k) + (k-1)\big]\, \tilde{u}_a + \sum_{a \in \mathcal{A}_o^-} \tilde{v}_a \\ &= \tilde{u}_0 + \sum_{a \in \mathcal{A}_o^-} \tilde{v}_a. \end{aligned}$$

The summands $\tilde{u}_0, \{\tilde{v}_a\}_{a \in \mathcal{A}_o^-}$ live on the distinct blocks $M_0, \{M_a\}_{a \in \mathcal{A}_o^-}$, so by orthogonality

$$\left\| w_0 S^0 + \sum_{a \in \mathcal{A}_o^-} w_a S^a \right\|^2 = \|\tilde{u}_0\|^2 + \sum_{a \in \mathcal{A}_o^-} \|\tilde{v}_a\|^2. \tag{15}$$

Sum of weighted squared norms. Applying orthogonality to (13) and (14) (using that $\tilde{u}_a$ does not appear in S^a),

$$\|S^0\|^2 = \|\tilde{u}_0\|^2 + \sum_{a \in \mathcal{A}_o^-} \|\tilde{u}_a\|^2, \qquad \|S^a\|^2 = \|\tilde{v}_a\|^2 + \|\tilde{u}_0\|^2 + \sum_{a' \neq a} \|\tilde{u}_{a'}\|^2.$$

We can then form the weighted combination and collect coefficients,

$$\begin{aligned} w_0 \|S^0\|^2 + \sum_{a \in \mathcal{A}_o^-} w_a \|S^a\|^2 &= [(1-k)+k]\|\tilde{u}_0\|^2 + \sum_{a \in \mathcal{A}_o^-} [(1-k)+(k-1)]\|\tilde{u}_a\|^2 + \sum_{a \in \mathcal{A}_o^-} \|\tilde{v}_a\|^2 \\ &= \|\tilde{u}_0\|^2 + \sum_{a \in \mathcal{A}_o^-} \|\tilde{v}_a\|^2. \end{aligned}$$

Comparing with (15) gives (12), and combined with the vanishing divergence term, $g_t(x) = 0$ for all $x \in \mathbb{R}^n$ and $t \in [0,1]$. □

Corollary 1.3. *Under the hypotheses of Proposition 7, FKC applied to the base composition (2) is exactly equivalent to running the naïve denoising SDE (7).*

Proof. By Proposition 7, $g_t \equiv 0$, hence $\bar{g}_t \equiv 0$. The log-weight update (9) then gives $\mathrm{d}\omega_t = 0$, so all log-weights remain at their initial value of 0 throughout the trajectory. Systematic resampling proportional to $\exp(\omega_t^{(k)})$ therefore samples uniformly across particles at every step, leaving the ensemble unchanged. Consequently, the resulting marginal coincides exactly with that of the naïve denoising SDE. □

G.6 Connection of Theorem 1 to importance-sampling effective sample size

The sensitivity factor in Theorem 1 is exactly the population quantity estimated by importance-sampling effective sample size. Given $X_1, \ldots, X_N \sim P^i$ and weights $w_n \propto P^w(X_n)/P^i(X_n)$, define empirical relative ESS by

$$\widehat{\mathrm{rESS}}_i = \frac{\left(\sum_{n=1}^N w_n\right)^2}{N \sum_{n=1}^N w_n^2}.$$

If $X_1, \ldots, X_N$ are i.i.d. from P^i and $\mathbb{E}_{P^i}[R_i^2] < \infty$, then almost surely,

$$\widehat{\mathrm{rESS}}_i \longrightarrow \frac{\mathbb{E}_{P^i}[R_i]^2}{\mathbb{E}_{P^i}[R_i^2]} = \frac{1}{1+\chi^2(P^w \| P^i)},$$

where $\mathbb{E}_{P^i}[R_i] = 1$. This connection between importance-sampling ESS and chi-squared divergence is standard, see e.g. S. Agapiou & Stuart (2017). Consequently,

$$\kappa_i = \mathrm{rESS}_i^{-1/2}.$$

Thus source-to-composition relative ESS estimates the density-ratio term controlling local error amplification in Theorem 1.

G.7 Importance ratios for a base composition

The required density ratios for computing relative ESS for a base composition simplify considerably. For samples from P^{a_1},

$$\begin{aligned}\frac{P^w(x)}{P^{a_1}(x)} &= \frac{1}{P^{a_1}(x)}\left(\frac{1}{Z^w}\frac{P^{a_1}(x)P^{a_2}(x)}{P^0(x)}\right)\\ &= \frac{1}{Z^w}\frac{P^{a_2}(x)}{P^0(x)}.\end{aligned}$$

Thus the P^{a_1} importance weights require only the ratio P^{a_2}/P^0. The expression for P^{a_2} follows similarly. For the base source,

$$\begin{aligned}\frac{P^w(x)}{P^0(x)} &= \frac{1}{P^0(x)}\left(\frac{1}{Z^w}\frac{P^{a_1}(x)P^{a_2}(x)}{P^0(x)}\right)\\ &= \frac{1}{Z^w}\frac{P^{a_1}(x)}{P^0(x)}\frac{P^{a_2}(x)}{P^0(x)}.\end{aligned}$$

ESS is invariant to multiplying all weights by a common positive constant, so we may disregard Z^w. Therefore, ESS can be estimated from only P^{a_1}/P^0 and P^{a_2}/P^0.

G.8 Estimating ESS ratios with a conditional diffusion model

Obtaining these likelihood ratios is still difficult, however. Here, we discuss how to approximate the required likelihood ratios using the conditional diffusion model itself rather than an auxiliary density-ratio estimator. To avoid confusion with empirical estimates in this section, we use P_θ to denote learned model quantities instead of $\widehat{P}$. Following the interpretation used by Li et al. (2023), let $\mathcal{E}_\theta(x, a)$ denote a variational lower bound on the conditional log likelihood for $a \in \{0, a_1, a_2\}$ induced by the diffusion model's forward noising process. We write $\widehat{\mathcal{E}}_\theta(x, a)$ for the Monte Carlo estimate of this quantity obtained by sampling timesteps and performing single-step denoising. We use paired differences

$$d_{a_1,0}(x) = \widehat{\mathcal{E}}_\theta(x, a_1) - \widehat{\mathcal{E}}_\theta(x, 0), \qquad d_{a_2,0}(x) = \widehat{\mathcal{E}}_\theta(x, a_2) - \widehat{\mathcal{E}}_\theta(x, 0)$$

as proxies for

$$\log \frac{P_\theta^{a_1}(x)}{P_\theta^0(x)}, \qquad \log \frac{P_\theta^{a_2}(x)}{P_\theta^0(x)}.$$

When estimating each difference, the same sampled diffusion timestep and forward-noise realization are reused across conditions. This follows the construction of Li et al. (2023), who show that differences between conditional denoising errors under matched noise have substantially lower variance than the individual errors. Because our use depends on the scale and tails of the likelihood ratios rather than only their ordering, we use the model-consistent VLB with the appropriate timestep weighting rather than the simplified uniformly weighted score used by Li et al. (2023) for classification.

Using N held-out samples $X_1, \ldots, X_N \sim P^i$ from each corresponding source distribution, the estimated log importance weights are

$$\ell_{a_1}(x) = d_{a_2,0}(x), \qquad \ell_{a_2}(x) = d_{a_1,0}(x), \qquad \ell_0(x) = d_{a_1,0}(x) + d_{a_2,0}(x).$$

The empirical relative ESS is then

$$\widehat{\mathrm{rESS}}_i = \frac{\left(\sum_n e^{\ell_i(X_n)}\right)^2}{N \sum_n e^{2\ell_i(X_n)}}.$$

For numerical stability, we evaluate this expression in log space.

Limitations. Density-ratio estimation remains difficult, particularly because ESS depends on a second moment and is therefore sensitive to errors in the tails of the distributions. Additionally, a VLB difference is not generally an exact log-likelihood ratio. If

$$G_c(x) = \log P_\theta^c(x) - \mathcal{E}_\theta(x, c)$$

denotes the conditional variational gap, then

$$\mathcal{E}_\theta(x, a_1) - \mathcal{E}_\theta(x, 0) = \log \frac{P_\theta^{a_1}(x)}{P_\theta^0(x)} - \left(G_{a_1}(x) - G_0(x)\right).$$

A gap independent of x is harmless for ESS because it multiplies every importance weight by the same constant, but x-dependent variation in the gap can distort the estimated ratios. Thus, the empirical metric for diffusion models is only a proxy for the true relative ESS term, unless we additionally assume variational gaps are x-independent.

G.9 Global stability under stronger pathwise assumptions

The same ESS-related sensitivity factor from Theorem 1 also appears in a finite-perturbation bound when overlap and error magnitude are controlled along an entire path between the source collections. Let P_s^i, $s \in [0, 1]$, be a differentiable path of positive normalized source densities, let

$$P_s^w(x) \propto \frac{P_s^{a_1}(x) P_s^{a_2}(x)}{P_s^0(x)},$$

and define the log-density velocities

$$h_{i,s}(x) = \partial_s \log P_s^i(x).$$

Proposition 8 (Pathwise global stability)**.** *Assume $s \mapsto P_s^w$ is absolutely continuous as an L^1-valued map, the displayed derivatives exist for almost every s, and the integrals below are finite. Then*

$$\mathrm{TV}(P_0^w, P_1^w) \le \sum_{i \in \{0, a_1, a_2\}} \int_0^1 \sqrt{1 + \chi^2(P_s^w \| P_s^i)}\, \|h_{i,s}\|_{L^2(P_s^i)}\, ds.$$

Consequently, if $\chi^2(P_s^w \| P_s^i) \le C$ for every i and s, and

$$L_i := \int_0^1 \|\partial_s \log P_s^i\|_{L^2(P_s^i)}\, ds,$$

then

$$\mathrm{TV}(P_0^w, P_1^w) \le \sqrt{1 + C}\, (L_0 + L_{a_1} + L_{a_2}).$$

Proof. Let

$$f_s(x) = \frac{P_s^{a_1}(x) P_s^{a_2}(x)}{P_s^0(x)}, \qquad Z_s^w = \int f_s(x)\, dx,$$

so that $P_s^w = f_s / Z_s^w$, and define $H_s = h_{a_1,s} + h_{a_2,s} - h_{0,s}$. Then

$$\partial_s \log f_s = H_s,$$

so $\partial_s f_s = f_s H_s$. Consequently,

$$\frac{\partial_s Z_s^w}{Z_s^w} = \frac{1}{Z_s^w} \int f_s(x) H_s(x)\, dx = \mathbb{E}_{P_s^w}[H_s].$$

Differentiating $P_s^w = f_s/Z_s^w$ yields

$$\partial_s P_s^w = P_s^w \left(H_s - \mathbb{E}_{P_s^w}[H_s]\right).$$

Therefore

$$\begin{aligned}\mathrm{TV}(P_0^w, P_1^w) &\le \frac{1}{2}\int_0^1 \mathbb{E}_{P_s^w}\left|H_s - \mathbb{E}_{P_s^w} H_s\right| ds \\ &\le \sum_i \int_0^1 \mathbb{E}_{P_s^w}|h_{i,s}|\, ds.\end{aligned}$$

For each term,

$$\mathbb{E}_{P_s^w}|h_{i,s}| = \mathbb{E}_{P_s^i}\left[\frac{P_s^w}{P_s^i}|h_{i,s}|\right] \le \sqrt{1+\chi^2(P_s^w\|P_s^i)}\,\|h_{i,s}\|_{L^2(P_s^i)},$$

which proves the first inequality. The uniform-C bound follows immediately. □

This result shows that the same ESS factor appearing in the local result from Theorem 1 also controls a global bound under stronger pathwise control assumptions.

As in the local bound, Proposition 8 extends directly to arbitrary weighted compositions $P_s^w \propto \prod_i (P_s^i)^{w_i}$. The argument proceeds identically after replacing $H_s = h_{a_1,s} + h_{a_2,s} - h_{0,s}$ with $H_s = \sum_i w_i h_{i,s}$. This yields

$$\mathrm{TV}(P_0^w, P_1^w) \le \sum_i |w_i| \int_0^1 \sqrt{1+\chi^2(P_s^w\|P_s^i)}\,\|h_{i,s}\|_{L^2(P_s^i)}\, ds.$$

G.10 A direct global bound from combined relative L^2 error

In this and the following subsection, we explore bounds obtained through stronger assumptions on the form of the errors.

The source-wise relative-L^2 assumptions from Theorem 1 are local. A global bound follows immediately if one instead controls the relative error of the entire unnormalized composition under the target distribution. Let

$$f = \frac{P^{a_1}P^{a_2}}{P^0}, \qquad \widehat{f} = \frac{\widehat{P}^{a_1}\widehat{P}^{a_2}}{\widehat{P}^0}, \qquad R = \frac{\widehat{f}}{f}.$$

Proposition 9 (Combined relative-L^2 stability)**.** *If*

$$\mathbb{E}_{P^w}[(R-1)^2] \le \eta^2, \qquad \eta < 1,$$

then

$$\chi^2(\widehat{P}^w\|P^w) \le \frac{\eta^2}{(1-\eta)^2}, \qquad \mathrm{TV}(P^w, \widehat{P}^w) \le \frac{\eta}{2(1-\eta)}.$$

Proof. Since $\widehat{P}^w/P^w = R/\mathbb{E}_{P^w}[R]$, Cauchy–Schwarz gives

$$\mathbb{E}_{P^w}[R] \ge 1 - \mathbb{E}_{P^w}|R-1| \ge 1-\eta.$$

Furthermore, $\mathrm{Var}_{P^w}(R) \le \mathbb{E}_{P^w}[(R-1)^2] \le \eta^2$, and hence

$$\chi^2(\widehat{P}^w\|P^w) = \mathbb{E}_{P^w}\left[\left(\frac{R}{\mathbb{E}_{P^w}[R]} - 1\right)^2\right] = \frac{\mathrm{Var}_{P^w}(R)}{\mathbb{E}_{P^w}[R]^2} \le \frac{\eta^2}{(1-\eta)^2}.$$

The TV bound follows from $\mathrm{TV}(P^w, \widehat{P}^w) \le \frac{1}{2}\sqrt{\chi^2(\widehat{P}^w\|P^w)}$. □

This proposition controls the error of the already-combined density and is therefore substantially stronger than separate relative-L^2 assumptions for each source as in Theorem 1. In particular, separate conditions $\chi^2(\widehat{P}^i\|P^i) \to 0$ do not prevent $\widehat{P}^0/P^0$ from approaching zero on a small set, where the reciprocal $P^0/\widehat{P}^0$ appearing in the composition may become arbitrarily large.

G.11 TV bound through uniform control of log-density error

If we are willing to make other strong assumptions on the form of the errors, namely uniform control of the log-density density error, we can obtain stronger bounds on the composed distributional error. First, assume the true and perturbed compositions are valid and the relevant density ratios are positive on the support of P^w. Write

$$\widehat{P}^i(x) = P^i(x)e^{\delta_i(x)}, \qquad \Delta(x) = \delta_{a_1}(x) + \delta_{a_2}(x) - \delta_0(x).$$

Then the unnormalized composition is multiplied by e^{Δ}, and normalization gives the identity

$$\frac{\widehat{P}^w(x)}{P^w(x)} = \frac{e^{\Delta(x)}}{\mathbb{E}_{P^w}[e^{\Delta}]}.$$

Thus only the variation of the combined log-density error Δ matters, since adding a constant to Δ has no effect on the normalized composition. More precisely, we have the following proposition.

Proposition 10 (Uniform control of combined log-density error)**.** *If*

$$\operatorname{osc}_{P^w}(\Delta) := \operatorname*{ess\,sup}_{P^w} \Delta - \operatorname*{ess\,inf}_{P^w} \Delta \le \eta,$$

then

$$\mathrm{TV}(P^w, \widehat{P}^w) \le \tanh(\eta/4).$$

Proof. Let $r = d\widehat{P}^w/dP^w = e^{\Delta}/\mathbb{E}_{P^w}[e^{\Delta}]$. Then $\mathbb{E}_{P^w}[r] = 1$ and the ratio between the essential supremum and infimum of r is at most e^{η}. If $a \le r \le b$ with $a \le 1 \le b$, convexity of $|r-1|$ implies

$$\mathrm{TV}(P^w, \widehat{P}^w) = \frac{1}{2}\mathbb{E}_{P^w}|r-1| \le \frac{(1-a)(b-1)}{b-a}.$$

Since $b/a \le e^{\eta}$, the right-hand side is maximized at $a = e^{-\eta/2}$ and $b = e^{\eta/2}$, yielding $\tanh(\eta/4)$. □

As an immediate corollary, we can also state a sufficient condition based on individual source errors. If

$$1 - \epsilon_i \le \frac{\widehat{P}^i(x)}{P^i(x)} \le 1 + \epsilon_i, \qquad \epsilon_i < 1,$$

then the unnormalized composition ratio $R = (\widehat{P}^{a_1}\widehat{P}^{a_2}/\widehat{P}^0)/(P^{a_1}P^{a_2}/P^0)$ satisfies $L \le R \le U$, where

$$L = \frac{(1-\epsilon_{a_1})(1-\epsilon_{a_2})}{1+\epsilon_0}, \qquad U = \frac{(1+\epsilon_{a_1})(1+\epsilon_{a_2})}{1-\epsilon_0},$$

and Proposition 10 gives

$$\mathrm{TV}(P^w, \widehat{P}^w) \le \frac{\sqrt{U}-\sqrt{L}}{\sqrt{U}+\sqrt{L}} = \tanh\left(\frac{1}{4}\log\frac{U}{L}\right).$$

However, these are extremely strong assumptions to make on the errors, and likely are not achievable in practice.

G.12 Stability conditions in Bayes composition

Because Bayes composition is a common construction for composing conditional diffusion models (Bradley et al., 2025; Liu et al., 2022; Du et al., 2023), we also analyze stability under the additional restriction that the source distributions arise as conditionals of a common joint distribution. As a preliminary, we call source distributions (P^0, P^{a_1}, P^{a_2}) ***Bayes-coherent*** if there exists a joint distribution P over (X, A, B) satisfying $A \perp B \mid X$ such that $P^0 = P_X$, $P^{a_1} = P(X \mid A)$, and $P^{a_2} = P(X \mid B)$. We similarly call a perturbation Bayes-coherent if the perturbed source triple arises from some perturbed joint distribution $\widehat{P}$ satisfying $A \perp B \mid X$. The ratio identities below require only that the true sources are Bayes-coherent, we require Bayes-coherence of the perturbation only when explicitly stated.

Let P be such a joint distribution and write

$$P^0(x) = p(x), \qquad P^{a_1}(x) = p(x \mid A), \qquad P^{a_2}(x) = p(x \mid B).$$

Then

$$P^w(x) = p(x \mid A, B) = \frac{1}{Z^w}\frac{P^{a_1}(x)P^{a_2}(x)}{P^0(x)}.$$

G.12.1 Bayes identities, overlap, and ESS

Writing $a(x) = P(A \mid X = x)$ and $b(x) = P(B \mid X = x)$, conditional independence gives

$$P^w(x) = \frac{a(x)b(x)P^0(x)}{P(A,B)}, \qquad Z^w = \frac{P(A,B)}{P(A)P(B)}.$$

Assuming the relevant conditional densities are positive wherever the ratios below are evaluated,

$$\frac{P^w(x)}{P^{a_1}(x)} = \frac{P(B \mid X = x, A)}{P(B \mid A)} = \frac{P(B \mid X = x)}{P(B \mid A)} \le \frac{1}{P(B \mid A)},$$

with a similar bound for P^w/P^{a_2}, while

$$\frac{P^w(x)}{P^0(x)} = \frac{P(A \mid X = x)P(B \mid X = x)}{P(A,B)} \le \frac{1}{P(A,B)}.$$

Thus, if $P(A,B) \ge \alpha > 0$, all three density ratios are bounded by $1/\alpha$.

These ratio bounds directly imply lower bounds on the corresponding ESS values. Since $R_i = P^w/P^i$, $R_i^2 \le \|R_i\|_\infty R_i$, and $\mathbb{E}_{P^i}[R_i] = 1$,

$$1 + \chi^2(P^w \| P^i) = \mathbb{E}_{P^i}[R_i^2] \le \|R_i\|_\infty.$$

Therefore

$$\mathrm{rESS}_{a_1} \ge P(B \mid A), \qquad \mathrm{rESS}_{a_2} \ge P(A \mid B), \qquad \mathrm{rESS}_0 \ge P(A,B).$$

Equivalently, the local sensitivity factors obey

$$\kappa_{a_1} \le P(B \mid A)^{-1/2}, \qquad \kappa_{a_2} \le P(A \mid B)^{-1/2}, \qquad \kappa_0 \le P(A,B)^{-1/2}.$$

Corollary 1.4 (Bayes-specific relative-L^2 sensitivity)**.** *Under the Bayes assumptions above, the perturbation from Theorem 1 satisfies*

$$\mathrm{TV}(P^w, P_s^w) \le s\left(\frac{\|u_{a_1}\|_{L^2(P^{a_1})}}{\sqrt{P(B \mid A)}} + \frac{\|u_{a_2}\|_{L^2(P^{a_2})}}{\sqrt{P(A \mid B)}} + \frac{\|u_0\|_{L^2(P^0)}}{\sqrt{P(A,B)}}\right) + o(s).$$

The corollary follows immediately by substituting the Bayes ratio bounds into Theorem 1. Proposition 9 also applies unchanged in the Bayes setting when the relative error of the combined unnormalized composition can be controlled directly under P^w.

Intuitively, these bounds state that if the joint conditioning event is probable enough, then the local sensitivity is bounded. However, this fact is not as useful as it may initially seem. In practice, for many compositional generation settings, we would like to generate samples from the composed distribution precisely because the target combination is one that is rare or absent in the available data. That is, the precise reason we would like to sample from the composition is also what makes such a composition potentially unstable.

G.12.2 Coherent perturbations of the underlying joint distribution

The Bayes structure yields a stronger stability statement when the perturbation is Bayes-coherent and the estimated source conditionals arise from a joint distribution $\hat{P}$ that is close to P rather than being perturbed independently. We note however that, for a learned model, this is a strong assumption on the form of the estimation error and is not generally guaranteed. As we will see in Sec. G.12.3, without this nearby-coherence assumption Bayes composition is no more immune to pathological error blowup than general base composition.

Proposition 11 (Stability of coherent conditioning)**.** *Let P and $\widehat{P}$ be joint distributions over (X, A, B) and let $E = \{A = 1, B = 1\}$. If*

$$\mathrm{TV}(P, \widehat{P}) \le \epsilon, \qquad P(E) \ge \alpha > 0, \qquad \epsilon < \alpha,$$

then $\widehat{P}(E) > 0$ and

$$\mathrm{TV}\Big(P(X \mid E), \widehat{P}(X \mid E)\Big) \le \frac{2\epsilon}{\alpha}.$$

Proof. Let $p = P(E)$ and $\widehat{p} = \widehat{P}(E)$. TV controls event probabilities, so $|p - \widehat{p}| \le \epsilon$ and therefore $\widehat{p} > 0$. For any measurable set C,

$$\begin{aligned}\left|P(C \mid E) - \widehat{P}(C \mid E)\right| &= \left|\frac{P(C \cap E)}{p} - \frac{\widehat{P}(C \cap E)}{\widehat{p}}\right| \\ &\le \frac{|P(C \cap E) - \widehat{P}(C \cap E)|}{p} + \widehat{P}(C \cap E)\left|\frac{1}{p} - \frac{1}{\widehat{p}}\right| \\ &\le \frac{\epsilon}{p} + \frac{|p - \widehat{p}|}{p} \\ &\le \frac{2\epsilon}{\alpha}.\end{aligned}$$

Taking the supremum over C proves the result. □

Proposition 11 is a general stability result for conditioning and does not itself require conditional independence. When both P and $\widehat{P}$ satisfy the Bayes conditional-independence structure, $P(X \mid A, B)$ and $\widehat{P}(X \mid A, B)$ coincide with the product-form base compositions. Thus, it yields stability of Bayes composition when the underlying joint distributions are close in TV, provided the joint conditioning event remains bounded away from zero.

G.12.3 Bayes valid compositions can still be sensitive to source errors

Even if the true sources arise from a valid Bayes model with a non-rare conditioning event, independently perturbing their densities can produce arbitrarily large error.

Proposition 12 (Instability under incoherent perturbations of Bayes-valid sources)**.** *Let $X \sim \mathrm{Unif}[0,1]$ and let A and B be independent of X and of each other, with $P(A) = P(B) = 1/2$. Then $P^0 = P^{a_1} = P^{a_2} = P^w = 1$ and $P(A, B) = 1/4$. For $E_\epsilon = [0, \epsilon]$, keep $\widehat{P}^{a_1} = \widehat{P}^{a_2} = 1$ and define*

$$\widehat{P}^0(x) = \begin{cases}\epsilon^2, & x \in E_\epsilon, \\ c_\epsilon, & x \notin E_\epsilon,\end{cases} \qquad c_\epsilon = \frac{1 - \epsilon^3}{1 - \epsilon}.$$

Then $\mathrm{TV}(P^0, \widehat{P}^0) = \epsilon - \epsilon^3 \to 0$ while the composition of the perturbed sources satisfies $\mathrm{TV}(P^w, \widehat{P}^w) \to 1$.

Proof. The choice of c_ϵ normalizes $\widehat{P}^0$. Since $\widehat{P}^0 < 1$ on E_ϵ and $\widehat{P}^0 > 1$ on its complement,

$$\mathrm{TV}(P^0, \widehat{P}^0) = \frac{1}{2}\left[\epsilon(1 - \epsilon^2) + (1 - \epsilon)(c_\epsilon - 1)\right] = \epsilon - \epsilon^3.$$

The perturbed unnormalized composition is $\widehat{f} = 1/\widehat{P}^0$. Its unnormalized mass on E_ϵ is

$$\int_{E_\epsilon} \widehat{f}(x)\, dx = \frac{1}{\epsilon},$$

whereas its mass on the complement is $(1 - \epsilon)/c_\epsilon = O(1)$. Hence $\widehat{P}^w(E_\epsilon) \to 1$, while $P^w(E_\epsilon) = \epsilon \to 0$, and therefore

$$\mathrm{TV}(P^w, \widehat{P}^w) \ge \left|\widehat{P}^w(E_\epsilon) - P^w(E_\epsilon)\right| \to 1.$$

□

Further, we can see that the perturbed distributions used above are Bayes-coherent. Let $q_\epsilon = \widehat{P}^0$, define $\widehat{P}_X = q_\epsilon$, and let $\widehat{P}(A = 1 \mid X = x) = \widehat{P}(B = 1 \mid X = x) = \frac{\epsilon^2}{q_\epsilon(x)}$, with $A \perp B \mid X$. Since $q_\epsilon(x) \ge \epsilon^2$, these are valid probabilities, and $\widehat{P}(A = 1) = \widehat{P}(B = 1) = \epsilon^2$. Bayes' rule then gives $\widehat{P}(X \mid A = 1) = \widehat{P}(X \mid B = 1) = 1$, which yields exactly the perturbed distributions used in the example. However, the joint distribution which the perturbations are coherent to is not close to P, since $\mathrm{TV}(P, \widehat{P}) \ge \left|P(A = 1) - \widehat{P}(A = 1)\right| = \frac{1}{2} - \epsilon^2$.

Consequently, we see that the stability statement of Proposition 11 applies only to coherent perturbations of the underlying joint model which arise from a *nearby* joint model. It is not in general true for Bayes compositions without making the strong assumption that perturbations induced by estimation error are coherent to a joint model, and in fact even this is not sufficient, we require the stronger assumption that they are coherent to a nearby joint model.

H Additional experimental results

H.1 Synthetic data: Gaussian mixture models

In a slightly more complex synthetic setting than the 2D Gaussians, we consider two GMM experiments that both violate the factorized conditionals assumption. In both experiments, $\mathcal{X} = \mathbb{R}^2$ and we define P^0, P^{a_1}, and P^{a_2} as mixtures of Gaussians, where each component has covariance matrix σI. Because P^w is a ratio of Gaussian mixtures, its normalizing constant is analytically intractable and the resulting density is not itself a GMM. We therefore obtain ground-truth samples via rejection sampling for in-distribution setting and importance sampling for out-of-distribution setting (I.2). Results are provided in Table 5.

For the in-distribution setting, component means are arranged on a 2×3 grid, $\mu_k \in \{-3, 0, 3\} \times \{-2.5, 2.5\}$, and we set $\sigma = 0.6$. The base P^0 assigns uniform weight to all six components. Conditions restrict to overlapping subsets: P^{a_1} is uniform over $\{k_0, k_1, k_2, k_3\}$, and P^{a_2} is uniform over $\{k_1, k_2, k_4, k_5\}$. This configuration is explicitly constructed to isolate approximation error: the unnormalized target density $P^{a_1}(x)P^{a_2}(x)/P^0(x)$ is confined to regions with high probability mass under the conditionals. By ensuring the composed target modes are well-represented in the training data, distributional estimation error is minimized. In this regime, FKC successfully corrects for the non-factorized nature of the composition, aligning the empirical distributions with the ground truth. Only for poorly learned models ($N = 100$) and at large particle counts does further increasing the number of particles begin to harm performance.

For the out-of-distribution setting, we design the modes so the target distribution concentrates at locations that are low-probability under all conditionals. P^{a_1} is a two-component GMM with means $\{(-6, 1.5), (9, 0)\}$, P^{a_2} has means $\{(0, -9), (-1.5, 6)\}$, and P^0 has all four means with uniform weight, all with $\sigma = 1$. While small increases in K yield initial improvements by correcting for approximation error, further scaling K worsens results due to accumulating estimation error; increasing the number of particles is only consistently helpful with oracle scores.

Table 5: **The expected trends also hold for Gaussian mixture models.** The table follows the same conventions as Table 1.

(a) **In-Distribution**

$K\backslash N$	100	1000	10000	analytic
1	1.2113	1.2470	1.1704	1.2133
4	0.6678	0.2842	0.2318	0.2414
16	0.7735	0.2206	0.1454	0.1044
64	0.8130	0.2134	0.1478	0.0736
256	0.8288	0.1985	0.1350	0.0771

(b) **Out-of-Distribution**

$K\backslash N$	100	1000	10000	analytic
1	4.0336	4.0508	3.9606	4.2102
4	2.7357	2.7265	2.4675	2.5081
16	2.4276	2.4556	2.1718	1.7704
64	2.6212	2.7205	2.4005	1.5531
256	3.0238	3.0562	2.6706	1.4670

H.2 Image generation: full accuracies and warping scores

Here we present full marginal and joint accuracies, as well as room warping scores, for image generation experiments. Experimental details for assessing image generation accuracy and warping scores are given in Apps. I.3 and I.4.

Table 6: **Compositional success in room generation follows similar trends to synthetic settings.** Judged results on 500 generated room images per setting (4000 images total), comparing naïve denoising ($K = 1$ particles) against FKC ($K = 16$). An image is scored by how frequently the given object is detected in the image by `gpt-5.6-luna`, even if warped. 'Joint' measures how frequently both objects are detected simultaneously (compositional success). An image warping score is also assigned by the same LLM; increasing the number of particles consistently increases the degree of warping.

(a) Factorized Conditional + ID

K	Joint	Couch	Painting	Warping
1	0.778	0.888	0.884	0.81
16	0.868	0.928	0.926	0.99

(b) Non-Factorized Conditional + ID

K	Joint	Couch	Table	Warping
1	0.648	0.900	0.710	1.37
16	0.792	0.932	0.838	1.36

(c) Factorized Conditional + OOD

K	Joint	Couch	Painting	Warping
1	0.406	0.674	0.726	0.87
16	0.250	0.604	0.626	1.04

(d) Non-Factorized Conditional + OOD

K	Joint	Couch	Table	Warping
1	0.142	0.784	0.296	1.57
16	0.098	0.742	0.282	1.62

Table 7: CelebA composition results for images generated from pixel-space diffusion models. Joint is the success rate at generating an image with both attributes as measured by FaceXFormer, the per-attribute columns are the marginal detection rates across 1000 images per cell (8000 total).

(a) Factorized Conditional + ID

K	Joint	Bangs	Mouth Slightly Open
1	0.979	0.990	0.989
16	0.991	0.991	1.000

(b) Non-Factorized Conditional + ID

K	Joint	Brown Hair	Wavy Hair
1	0.879	0.924	0.947
16	0.951	0.958	0.993

(c) Factorized Conditional + OOD

K	Joint	Bangs	Mouth Slightly Open
1	0.746	0.824	0.904
16	0.642	0.810	0.798

(d) Non-Factorized Conditional + OOD

K	Joint	Brown Hair	Wavy Hair
1	0.523	0.734	0.687
16	0.517	0.773	0.626

H.3 Latent diffusion models

In addition to pixel-space diffusion models, we also run experiments with latent diffusion models (Rombach et al., 2022; Podell et al., 2024). For each setting, we train an encoder to embed images, and diffusion is performed purely in its latent space. Details are given in App. I.5.

Table 8: Room composition results with images generated from a latent diffusion model. Setup follows Table 6.

(a) Factorized Conditional + ID

K	Joint	Couch	Painting	Warping
1	0.786	0.890	0.894	0.61
16	0.732	0.790	0.930	0.53

(b) Non-Factorized Conditional + ID

K	Joint	Couch	Table	Warping
1	0.706	0.896	0.778	1.15
16	0.666	0.860	0.740	1.20

(c) Factorized Conditional + OOD

K	Joint	Couch	Painting	Warping
1	0.430	0.658	0.770	0.75
16	0.304	0.508	0.792	0.63

(d) Non-Factorized Conditional + OOD

K	Joint	Couch	Table	Warping
1	0.116	0.742	0.332	1.22
16	0.032	0.478	0.438	1.21

In Table 8, we observe that increasing the particle count worsens performance in every quadrant. This effect is strongest in OOD settings, following previous trends. To diagnose why this occurs, we perform an experiment where we generate with the same model, but halt FKC resampling within the last 10 (of 100) steps. This corresponds to only resampling within an active region, as proposed in Skreta et al. (2025). Under this alteration, we observe performance to improve across the board, and FKC once again improves accuracy in ID settings, though only marginally (Table 9). Since $K = 1$ involves only a single particle, it is not affected by this change, and we observe the judge to provide very similar judgements as before.

Table 9: Room composition results with images generated from a latent diffusion model and FKC gated after step 90. Setup otherwise follows Table 8.

(a) Factorized Conditional + ID

K	Joint	Couch	Painting	Warping
1	0.790	0.894	0.894	0.59
16	0.810	0.894	0.910	0.67

(b) Non-Factorized Conditional + ID

K	Joint	Couch	Table	Warping
1	0.708	0.894	0.782	1.13
16	0.712	0.904	0.794	1.16

(c) Factorized Conditional + OOD

K	Joint	Couch	Painting	Warping
1	0.430	0.662	0.764	0.72
16	0.362	0.564	0.798	0.75

(d) Non-Factorized Conditional + OOD

K	Joint	Couch	Table	Warping
1	0.134	0.754	0.350	1.22
16	0.054	0.538	0.432	1.23

Table 10: CelebA composition results for images generated from latent diffusion models. Setup follows Table 7.

(a) Factorized Conditional + ID

K	Joint	Bangs	Mouth Slightly Open
1	0.967	0.986	0.981
16	0.981	0.981	1.000

(b) Non-Factorized Conditional + ID

K	Joint	Brown Hair	Wavy Hair
1	0.865	0.929	0.920
16	0.886	0.932	0.944

(c) Factorized Conditional + OOD

K	Joint	Bangs	Mouth Slightly Open
1	0.678	0.808	0.834
16	0.494	0.732	0.707

(d) Non-Factorized Conditional + OOD

K	Joint	Brown Hair	Wavy Hair
1	0.499	0.728	0.676
16	0.448	0.786	0.532

H.4 Separate diffusion models

Instead of training a single diffusion model and composing its conditional distributions, it is possible to use separate diffusion models for the conditions, and compose in the same fashion via their product. This is desirable since, for instance, pre-trained diffusion models may possess different capabilities, and composing existing models may allow for greater control or improved performance on particular tasks than individually using any one pre-trained model.

To consider this approach, we re-run our synthetic experimental settings, including the room dataset, individually training a separate diffusion model for each conditional distribution. Full results from these experiments are in K.2 and K.4.

We observe that despite learning the underlying conditional distributions to the same level of in-distribution accuracy as conditional models, the compositions of these individual experts can perform worse in practice than their conditional model counterpart, as Table 11 demonstrates for 2D Gaussians. This evidence suggests that implicit regularization through the weight sharing of the conditional diffusion model may play an important role in mitigating out-of-distribution distributional estimation error and enabling effective weighted compositions.

Table 11: **The expected trends also hold for individual "expert" models, but overall performance is slightly worse and OOD estimation error accumulates more quickly.** The table follows the same conventions and set up as Table 1, but with individual models in place of a single conditional model.

(a) Factorized conditionals + ID

$K\backslash N$	100	1000	10000	analytic
1	0.4443	0.1339	0.0933	0.0370
4	0.5546	0.1217	0.0765	0.0363
16	0.6076	0.1177	0.0763	0.0372
64	0.6267	0.1140	0.0743	0.0365
256	0.6364	0.1196	0.0786	0.0359

(b) Non-factorized conditionals + ID

$K\backslash N$	100	1000	10000	analytic
1	0.5553	0.1437	0.1106	0.0673
4	0.6524	0.1213	0.0781	0.0389
16	0.6930	0.1148	0.0711	0.0356
64	0.7034	0.1147	0.0705	0.0358
256	0.7110	0.1157	0.0737	0.0350

(c) Factorized conditionals + OOD

$K\backslash N$	100	1000	10000	analytic
1	1.1771	0.8397	0.7643	0.1163
4	2.2327	2.0231	2.0465	0.1160
16	3.1107	3.2394	3.4382	0.1174
64	3.5605	3.9171	4.2961	0.1143
256	3.8044	4.2434	4.7380	0.1061

(d) Non-factorized conditionals + OOD

$K\backslash N$	100	1000	10000	analytic
1	1.2463	1.0431	0.8929	0.6418
4	2.3105	2.2133	1.8170	0.1704
16	3.1794	3.4853	2.9968	0.0905
64	3.6653	4.2021	3.8652	0.0821
256	3.8937	4.5325	4.4165	0.0815

Table 12: Room composition results for images generated from composing "expert" diffusion models individually trained for each condition.

(a) Factorized Conditional + ID

K	Joint	Couch	Painting	Warping
1	0.802	0.890	0.894	1.22
16	0.924	0.972	0.944	1.45

(b) Non-Factorized Conditional + ID

K	Joint	Couch	Table	Warping
1	0.774	0.956	0.796	1.64
16	0.956	0.992	0.960	1.83

(c) Factorized Conditional + OOD

K	Joint	Couch	Painting	Warping
1	0.274	0.638	0.630	1.26
16	0.178	0.636	0.538	1.43

(d) Non-Factorized Conditional + OOD

K	Joint	Couch	Table	Warping
1	0.226	0.820	0.338	1.72
16	0.214	0.850	0.296	1.95

However, Table 12 demonstrates that, in room image generation, separate models can improve performance within in-distribution settings. Yet, this also comes with a pronounced decrease in compositionality in the factorized OOD setting. Further, composing separate models substantially increases warping scores across the board. This may, in part, explain the apparent increase in compositional success: in qualitative analysis, the LLM judge appears to be especially lenient with warped furniture, and sometimes claims an object is present even if it is warped beyond recognition.

H.5 Sensitivity metric in image generation

Since $\kappa_i = \mathrm{rESS}_i^{-1/2}$, $-\log \mathrm{rESS}_i = 2 \log \kappa_i$ is a monotone measure of sensitivity. We report these logarithm metrics, as they place numbers in a more interpretable range. For the source errors, the local bound depends on $\sum_i \kappa_i \|u_i\|_{L^2(P^i)}$. To obtain a single scalar sensitivity metric for multiplying the total source error $\sum_i \|u_i\|_{L^2(P^i)}$, we have that $\max_i \kappa_i$ gives a valid upper bound, which is used in Figure 4. However, in our image generation experiments we empirically find that excluding the base distribution from the maximum results in more reliable separation. In particular, we find that $\max_{i \in \{0, a_1, a_2\}} - \log \mathrm{rESS}_i = - \min_{i \in \{0, a_1, a_2\}} \log \mathrm{rESS}_i$ is not consistently effective at separating ID settings from OOD settings for facial generation. Consequently, we use $\max(-\log \mathrm{rESS}_{a_1}, -\log \mathrm{rESS}_{a_2}) = \max_{i \in \{a_1, a_2\}} - \log \mathrm{rESS}_i$ for both room and face generation experiments to remain consistent.

We further find that it is important to measure this metric only with an active range. In particular, including timesteps t late in the denoising process leads to very high noise and renders the metric uninformative. While including timesteps t early in the denoising process is not harmful, we also find that these early timesteps are oftentimes less informative than later ones. As a result, we report results where sensitivity is estimated on both an unconstrained range, and a range with $t \in [100, 900]$.

For sake of space, we abbreviate Factorized Conditional as FC and Non-Factorized Conditional as NFC in this section.

Table 13: Room generation sensitivity results for $t \in [100, 900]$.

Setting	$-\log \mathrm{rESS}_0$	$-\log \mathrm{rESS}_{a_1}$	$-\log \mathrm{rESS}_{a_2}$	$\max_{i\in\{a_1,a_2\}} -\log \mathrm{rESS}_i$
FC ID	5.11	2.27	3.69	3.69
FC OOD	5.56	6.15	6.82	6.82
NFC ID	3.93	3.06	2.64	3.06
NFC OOD	6.88	6.93	5.57	6.93

Table 14: Room generation sensitivity results with the unconstrained range.

Setting	$-\log \mathrm{rESS}_0$	$-\log \mathrm{rESS}_{a_1}$	$-\log \mathrm{rESS}_{a_2}$	$\max_{i\in\{a_1,a_2\}} -\log \mathrm{rESS}_i$
FC ID	4.86	3.47	4.92	4.92
FC OOD	6.88	6.75	6.85	6.85
NFC ID	6.24	5.74	6.85	6.85
NFC OOD	6.93	6.15	6.90	6.90

Table 15: CelebA sensitivity results for $t \in [100, 900]$.

Setting	$-\log \mathrm{rESS}_0$	$-\log \mathrm{rESS}_{a_1}$	$-\log \mathrm{rESS}_{a_2}$	$\max_{i\in\{a_1,a_2\}} -\log \mathrm{rESS}_i$
FC ID	6.62	3.19	3.18	3.19
FC OOD	6.45	4.01	5.51	5.51
NFC ID	5.49	3.07	2.78	3.07
NFC OOD	4.65	4.13	3.47	4.13

Table 16: CelebA sensitivity results with the unconstrained range.

Setting	$-\log \mathrm{rESS}_0$	$-\log \mathrm{rESS}_{a_1}$	$-\log \mathrm{rESS}_{a_2}$	$\max_{i\in\{a_1,a_2\}} -\log \mathrm{rESS}_i$
FC ID	6.67	4.14	4.27	4.27
FC OOD	6.56	5.41	4.95	5.41
NFC ID	5.81	3.94	3.41	3.94
NFC OOD	5.36	4.89	3.56	4.89

Table 17: ID/OOD gap in $\max_{i\in\{a_1,a_2\}} -\log \mathrm{rESS}_i$, computed as OOD minus ID. Positive values indicate greater estimated sensitivity for the OOD composition.

Range	Rooms FC	Rooms NFC	CelebA FC	CelebA NFC
$t \in [100, 900]$	+3.13	+3.87	+2.32	+1.07
Unconstrained	+1.93	+0.05	+1.14	+0.94

I Experiment setup details

In this section, we provide additional experimental details. All experimental code is additionally provided at anonymous.4open.science/r/compose-anonymous-A3E1/. Code is primarily ran on the GPU, with either NVidia L40S or A6000s. The most expensive training code, which is for the image generation models, can be ran on a single GPU in around 9 hours. For each table, training and sampling across all configurations and across 30 runs consumes roughly 7 GPU hours. GPU-based computation is the primary computational bottleneck, memory and cpu requirements are minimal in comparison, though about 48GB of GPU memory is recommended for best performance.

I.1 2D Gaussian

Architecture. For each distribution we train a noise-prediction network $\epsilon_\theta(t, x)$ implemented as a 4-layer MLP with hidden width 512 and SiLU activations, predicting $\epsilon \in \mathbb{R}^2$ from the concatenation $[t, x]$. In the conditional setting a single network shared across $\{P^{a_1}, P^{a_2}, P^0\}$ takes $[t, x, \mathrm{onehot}(c)]$ as input, where $c \in \{0, 1, 2\}$ identifies the source distribution. For separate models, c is omitted. The score is recovered as $s(t, x) = -\epsilon_\theta(t, x)/\sigma(t)$. We note that these architectures are highly overparameterized for the particular learning task, consistent with many modern setups.

Training objective. We use denoising score matching in ϵ-prediction form (Ho et al., 2020).

Optimizer and schedule. Adam (Kingma & Ba, 2014) with learning rate 2×10^{-4}, default $(\beta_1, \beta_2) = (0.9, 0.999)$, $\varepsilon = 10^{-8}$, no weight decay, no learning-rate schedule. Batch size is set to 512. Training is run for 20,000 iterations in the single-shot pipeline and 20,000 iterations per cell in the sweep, a dataset of N points is created before training begins and sampled with replacement during training.

Noising schedule. Variance-preserving (VP) SDE (Song et al., 2021) with linear $\beta(\tau) = \beta_{\min} + \tau(\beta_{\max} - \beta_{\min})$, $\beta_{\min} = 0.1$, $\beta_{\max} = 20$, on the forward time $\tau \in [0, 1]$. (Note that for implementation cleanliness, we use the opposite convention as the rest of our paper: $\tau = 1 - t$ so that $t = 0$ is pure noise and $t = 1$ is clean data.)

Sampling. We implement the Feynman–Kac corrector sampler (Skreta et al., 2025) running N independent swarms of K particles in parallel, performs systematic resampling within every swarm at every step, and at the final step draw one sample from each swarm's weighted ensemble. We use 500 integration steps. We report metrics on $N = 5000$ output samples per configuration.

Sliced Wasserstein-2. Sliced $\mathcal{W}_2$ is computed via the Python Optimal Transport package (Flamary et al., 2021; 2024), with $p = 2$ and 2000 random projections per evaluation.

Maximum Mean Discrepancy. We report the unbiased MMD^2 U-statistic with the Gaussian RBF kernel $k(x, y) = \exp\big(-\gamma\|x - y\|_2^2\big)$. γ is set by the median heuristic on the empirical cross-pairwise Euclidean distances between the two sample sets, with $\gamma = 1/(2\mathrm{med}^2)$.

Reference samples. Because the target is a closed-form Gaussian, we draw ground-truth samples directly via its diagonal covariance. Sliced $\mathcal{W}_2$ and MMD^2 both use 5000 ground-truth samples per evaluation.

Seeds. The base seed is 1 and is applied at the start of each run. For multiple runs N, the run r $(0 \le r < N)$ uses seed $1 + r$, which seeds every random component of the process. We average over $n_{\text{runs}} = 30$ independent runs and report mean $\pm$ standard deviation.

Clipping. During the FKC denoisnig process, we introduce an optional `g_clip` paramater, which clips the $g_t(x)$ increment (10) at every step to $[-\texttt{g_clip}, \texttt{g_clip}]$. Clipping generally reduces distributional estimation error by preventing weights from concentrating on erroneously high values, but can prevent score approximation error from decreasing as quickly as it would otherwise, particularly for oracle or well-learned models. In practice, we find this approach to be an effective method of trading off estimation error for approximation error. For 2D Gaussian and Mixture of Gaussians experiments, we set $\texttt{g_clip} = 15.0$. We omit clipping for image generation, as we observe high variability in the magnitude of the g_t updates. Late in the denoising process, updates are orders of magnitude larger than those earlier on, so an effective clip parameter is difficult to tune, which is not the case for the synthetic experiments. Qualitatively, we observed strong clipping to reduce compositionality, but also reduce smearing/warping.

I.2 Mixture of Gaussians

Unless otherwise explicitly noted, all neural network architectures, noising schedules, and optimization hyperparameters for the GMM experiments are identical to those described for the 2D Gaussian setting in App. I.1.

Rejection sampling (In-Distribution) To obtain exact samples from the target distribution $P^w(x) \propto \frac{P^{a_1}(x)P^{a_2}(x)}{P^0(x)}$, we utilize an exact rejection sampling scheme. We designate $P^{a_1}(x)$ as the proposal distribution, while $P^{a_2}(x)$ and $P^0(x)$ serve as the numerator ratio factor and the denominator, respectively. Mathematically, the modes of P^{a_2} are a subset of the modes of P^0. To compute the rejection bound analytically, we structure their implementations to share the exact same underlying GMM components (identical means and isotropic variances), where P^{a_2} assigns a mixture weight of zero to any mode outside its subset. Because $P^{a_2}(x)$ and $P^0(x)$ now differ strictly in their mixture weight arrays (denoted as w_{a_2} and w_0), the tight upper bound $M = \sup_x \frac{P^{a_2}(x)}{P^0(x)}$ can be computed exactly as $M = \max_k \frac{w_{a_2}[k]}{w_0[k]}$. Candidate samples are drawn from the proposal $x \sim P^{a_1}(x)$ and are subsequently accepted with probability $\frac{P^{a_2}(x)}{M \cdot P^0(x)}$.

Importance Sampling (Out-of-Distribution) In the out-of-distribution settings, the constituent GMMs do not share component means, rendering the analytical rejection bound intractable. For these cases, we employ Importance Sampling to obtain asymptotically exact samples. We compute the analytical product of all numerator GMMs to serve as the proposal distribution. We draw a heavily oversampled batch of candidate points from this proposal ($1000\times$ the target sample size) and assign each point a log-importance weight proportional to $-\sum_j \log p_j(x)$, representing the inverse effect of the denominator distributions. Finally, we perform systematic resampling based on these weights to extract n unweighted samples approximating the target distribution.

I.3 Objects in a room

For the sake of reproducibility, we produce full details here for our dataset construction as well as our model training and sampling methodologies.

Evaluation: object detection and warping scores. For the room experiments, we determine image presence through a VLM. Images are graded by GPT-5.6-LUNA at image detail `high` and reasoning effort `none`, with structured outputs over the boolean fields `couch_present` and `target_present` (boolean), as well as a field for scoring how warped/smeared the image is, `artifact_score` (integer 0–3), and a free-text rationale for the provided scores. The system prompt is "You are a consistent image-evaluation judge. Follow the provided rubric literally and return only the requested schema." The prompt is as follows, with `{target}` replaced by the condition's target object:

```
Judge this generated room image on three independent fields.
Object presence:
- couch_present:  true only if a visible object is
identifiable as a couch/sofa from its shape and structure.
Minor distortion is acceptable.
- target_present:  true only if a visible object is
identifiable as a {target} from its shape and structure.
Minor distortion is acceptable.
- Do not infer presence from the expected room contents,
placement alone, or a vague object-like blob.  If the object
is severely warped, merged with another object, or too
ambiguous to identify confidently, mark it false.
Overall image warping/artifacting:
- 0:  fully coherent; no visible generation-related
deformation, melting, duplication, broken geometry, or
meaningful texture artifacts.
- 1:  mild; one or a few visible but localized flaws.
Objects and room geometry remain structurally coherent.
- 2:  moderate; multiple obvious flaws, or one major
malformed/merged object or structural inconsistency, while
the overall scene remains understandable.
- 3:  severe; widespread or highly disruptive corruption,
badly incoherent room geometry, or major objects that are
substantially melted or merged.
Assess artifact_score across the whole image, including room
geometry and all objects.  When between two artifact scores,
choose the higher score.  Give a short factual rationale (25
words or fewer).
```

The prompt is followed by four human-labelled images for calibration purposes. Judging is performed for 500 images per setting and number of FKC particles, for 4000 images total (8000, including latent diffusion models). We report the per-object and joint presence rates, as well as the mean artifact score, for each group.

Dataset construction. To create our dataset of 256×256 images, we prompt the open-source text-to-image model FLUX.1-SCHNELL (Labs, 2024; Labs et al., 2025) with a set of custom prompts. The prompt for generating an empty room is:

> A photograph of an empty living room with plain white walls and wooden floors. The room has a large window, and it is sunny outside. The room is completely empty. It contains no furniture, no decorations, no plants, and no other objects. Completely undecorated. Abandoned but clean. The photo is wide angle, showing the entire room and how it is empty.

Class-conditioned prompts form variations on this, replacing the line

> The room is completely empty.

as follows:

- **couch:** The room is completely empty, except for a (couch:1.4)
- **coffee table:** The room is completely empty, except for a (coffee table:1.4)

- **framed painting:** The room is completely empty, except for a (framed painting:1.4) on the wall.
- **couch + coffee table**: The room is completely empty, except for a (couch:1.4) and a (coffee table:1.4).
- **couch + framed painting**: The room is completely empty, except for a (couch:1.4) and a (framed painting:1.4) on the wall.

Note that the numbers inside this prompt are interpreted by FLUX.1-SCHNELL as token weights and not as tokens, affording us more control over its prompt adherence.

After generating a set of candidate images, we manually label 1000 images from each class as either `accept` if they meet the prompt's description and do not suffer from artifacting, and `reject` otherwise. We allow small incidental objects built-in to the room, such as radiators or small ceiling lights.

We embed the labeled images with DINOv2-LARGE (Oquab et al., 2024). These embeddings are used to train a separate binary logistic regression classifier for each class to automatically accept or reject images. These classifiers are ran on the remaining samples (generating more as necessary) until we achieve 2000 accepted images per class, which then forms our dataset.

Architecture. Class-conditional 2-D U-Net (Ronneberger et al., 2015) (from the `diffusers` library (von Platen et al., 2022)) with input/output channels 3, sample size 256, two residual blocks per level, channel multipliers $(128, 256, 256, 512)$, down-blocks `(Down, Down, AttnDown, Down)`, up-blocks `(Up, AttnUp, Up, Up)`, and a learned class embedding indexed by the class (or condition) integer. The total number of class embeddings equals the number of training labels for each task, i.e. 4.

Forward process. Discrete DDPM (Ho et al., 2020) with $T = 1000$ training timesteps and a linear β schedule (Nichol & Dhariwal, 2021) via `diffusers.DDPMScheduler`. We use the v-prediction parametrization (Salimans & Ho, 2022), along with with terminal-SNR rescaling (Lin et al., 2024) to remove all residual signal at the terminal noise level. To acquire scores from v-prediction to perform FKC sampling, we convert v_t^a to a score via $S_t^a = -\epsilon_t^a / \sigma(t)$ with $\epsilon_t^a = \sigma(t)\, x_t + \alpha(t)\, v_t^a$. We use 100 integration steps during inference time.

Optimizer and schedule. AdamW (Loshchilov & Hutter, 2019) with learning rate 1×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\varepsilon = 10^{-8}$, no weight decay. Cosine learning-rate schedule with linear warmup over 500 steps into epochs $\times$ steps_per_epoch total updates. Mixed-precision training in `bfloat16`. An exponential moving average of the model weights with decay 0.9999 is maintained throughout training; sampling uses the EMA copy.

Data pipeline and hyperparameters. Per-image transform: `RandomHorizontalFlip`, `ToTensor`,`Normalize(mean, std)` where (mean, std) is the fixed $(0.5, 0.5, 0.5)$ pair. We balance classes during training according to the mixture probabilities of each condition. Batch size 16, 50 epochs.

I.4 FACIAL IMAGE GENERATION

This section contains the experimental details for the facial image generation (CelebA) experiments.

Evaluation: attribute detection. Generated CelebA images are scored with FaceXFormer (Narayan et al., 2025), using its attribute-prediction head. This outputs a sigmoid probability for each of the 40 CelebA attributes, we count each attribute of interest as present when its probability is at least 0.5. Each diffusion model generates and is evaluated on 1000 images per FKC particle count, resulting in 8000 images total per setting (16000 including the diffusion models).

Architecture. Before processing by the network, images are rescaled to 64×64. We use a 14.9M-parameter DDPM-style conditional U-Net (Ho et al., 2020), with base width 64 with channel multipliers $(1, 2, 2, 4)$, two residual blocks per level, GroupNorm and SiLU, dropout 0.1, and four-head

self-attention only at the 8×8 resolution. We encode the timestep as a 64-dimensional sinusoidal embedding, which a two-layer MLP projects to 256 dimensions, and it is summed with a learned 256-dimensional embedding of the condition $\{0, a_1, a_2\}$. (No $a_1 a_2$ condition is provided.) This vector uses adaptive group normalisation (Dhariwal & Nichol, 2021). The network predicts ϵ and is trained with an MSE loss.

Forward process, optimizer, and schedule. Discrete DDPM (Ho et al., 2020) with $T = 1000$ timesteps and a linear β schedule from 10^{-4} to 2×10^{-2}, in ϵ-prediction form. We use AdamW (Loshchilov & Hutter, 2019) with learning rate 2×10^{-4} and weight decay 10^{-4}. Learning rate has 5000 linear warmup steps followed by cosine decay, batch size is 64, and we use 200,000 optimiser steps. We use an exponential moving average of the weights with decay 0.9999; sampling uses this EMA copy and 100 inference steps. Images are augmented by augmented by a random horizontal flip.

I.5 Latent diffusion models

Both image-generation experiments are repeated with latent diffusion models Rombach et al. (2022). Each of the four cells per setting has its own autoencoder trained on that cell's training mixture, and its own diffusion model. The data mixtures are identical to their pixel-space counterparts. For room experiments the latent dimension is $2 \times 32 \times 32 = 2048$ versus $3 \times 256 \times 256 = 196{,}608$ in pixel space, for CelebA it is $3 \times 16 \times 16 = 768$ versus $3 \times 64 \times 64 = 12{,}288$.

Autoencoder. A KL-regularised autoencoder (`diffusers AutoencoderKL` (von Platen et al., 2022)) following the setup of the Stable Diffusion autoencoder (Rombach et al., 2022). We change the latent width and, for CelebA, the depth. The downsample factor is 8 ($256 \to 32$) with 2 latent channels for rooms, and 4 ($64 \to 16$, with one down-block removed) with 3 latent channels for CelebA. We use AdamW with learning rate 10^{-4}, $(\beta_1, \beta_2) = (0.5, 0.9)$, a cosine schedule with 500 warmup steps, batch size 16, 50 epochs, and an EMA of the weights with decay 0.9999. Diffusion training loads the EMA copy with the autoencoder frozen.

Latent standardization. Latents are standardized per channel by the mean and standard deviation of sampled posterior latents. This is computed once at the start of diffusion training over the training set.

Diffusion model. For both settings the diffusion model is largely identical to the pixel-space room model of App. I.3. The differences are that it operates on the standardized latents from the encoder, with input and output channels equal to the latent width. It is a three-level U-Net with channel multipliers $(128, 256, 512)$, down-blocks `(Down, AttnDown, Down)` and up-blocks `(Up, AttnUp, Up)`.

I.6 Sensitivity metric experiments

Here, we describe how we estimate the calculate the effective sample size (rESS) quantities from Sec. 4.3 and App. G.6.

Metric. For each source $i \in \{0, a_1, a_2\}$, we draw held-out samples $X_1, \ldots, X_N \sim P^i$ and form importance weights $w_n \propto P^w(X_n)/P^i(X_n)$. We estimate

$$\widehat{\mathrm{rESS}}_i = \frac{\left(\sum_{n=1}^N w_n\right)^2}{N \sum_{n=1}^N w_n^2}.$$

For the base composition, $w \propto P^{a_2}/P^0$ on samples from P^{a_1}, P^{a_1}/P^0 on samples from P^{a_2}, and $P^{a_1} P^{a_2}/(P^0)^2$ on samples from P^0. The unknown normalizing constant of P^w cancels. At the population level,

$$\kappa_i = \mathrm{rESS}_i^{-1/2}, \qquad -\log \mathrm{rESS}_i = 2 \log \kappa_i = \log\left(1 + \chi^2(P^w \| P^i)\right).$$

To present a single scalar metric, we aggregate across sources by

$$\mathrm{rESS} = \min_{i\in\{0,a_1,a_2\}} \mathrm{rESS}_i, \qquad -\log \mathrm{rESS} = \max_{i\in\{0,a_1,a_2\}} (-\log \mathrm{rESS}_i)\,.$$

This corresponds to the (slightly looser) worst-source sensitivity, since $\mathrm{rESS}^{-1/2} = \max_i \kappa_i$ gives a valid scalar value multiplying the total source error $\sum_i \|u_i\|_{L^2(P^i)}$ in the local bound.

Gaussian families. There are 8 families of 2D Gaussians with 20 configurations each, for 160 in total. Every family has base $P^0 = \mathcal{N}(0, I)$. Below, we give the source distributions, resulting composition, and the 20 parameter values for each family. For Gaussian distributions, the population $\mathbb{E}_{P^i}[w^2] = \infty$ if and only if $\Sigma^w \prec 2\,\Sigma^i$.

- **Anisotropic sources.** $P^{a_1} = \mathcal{N}(0, \mathrm{diag}(r, 1))$, $P^{a_2} = \mathcal{N}(0, \mathrm{diag}(1, r))$, so $P^w = \mathcal{N}(0, rI)$. The 20 values of r are evenly spaced in $[1.05, 2.10]$. The sources are widened along different axes, causing their precisions to cancel on each axis and leaving P^w isotropic with the variance of each source's wider axis. The population $\mathbb{E}_{P^i}[w^2] = \infty$ for the two configurations with $r > 2$.
- **Anisotropic sources, $r > 2$.** The same as above, with 20 values of r spaced evenly on a logarithmic scale in $[2, 200]$. The population $\mathbb{E}_{P^i}[w^2] = \infty$ throughout.
- **Orthogonal shifts.** $P^{a_1} = \mathcal{N}((m, 0), I)$, $P^{a_2} = \mathcal{N}((0, m), I)$, so $P^w = \mathcal{N}((m, m), I)$. The 20 values of m are evenly spaced in $[0.3, 3]$. The sources are shifted along orthogonal axes, while P^w is shifted along both.
- **Opposing shifts.** $P^{a_1} = \mathcal{N}((m, 0), I)$, $P^{a_2} = \mathcal{N}((-m, 0), I)$, so $P^w = \mathcal{N}(0, I)$. The 20 values of m are evenly spaced in $[0.5, 6]$. The source shifts cancel, so $P^w = P^0$ for every m.
- **Shared precision, $p < 1$.** $P^{a_1} = P^{a_2} = \mathcal{N}(0, I/p)$, so $P^w = \mathcal{N}(0, I/(2p-1))$. The 20 values of p are evenly spaced from 0.95 to 0.73. The identical sources are wider than the base, and P^w is even wider.
- **Shared precision, $p > 1$.** The same as above, with 20 values of p spaced evenly on a logarithmic scale in $[1.1, 16]$. The sources are narrower than the base, and P^w is even narrower.
- **P^w inside distant sources.** $P^{a_1} = P^{a_2} = \mathcal{N}((3, 3), s^2 I)$, so $P^w = \mathcal{N}\big(\frac{6}{2-s^2}(1, 1), \frac{s^2}{2-s^2} I\big)$. The 20 values of s are spaced evenly on a logarithmic scale from 0.6 to 0.03. The sources are narrow and far from the base, while P^w is narrower than the sources and lies within them.
- **Opposed anisotropic sources.** $P^{a_1} = \mathcal{N}(m(1/\sqrt{2}, -1/\sqrt{2}), \mathrm{diag}(1.35, 1.05))$, $P^{a_2} = \mathcal{N}(-m(1/\sqrt{2}, -1/\sqrt{2}), \mathrm{diag}(1.05, 1.35))$, so $P^w = \mathcal{N}(-0.305\,m\,(1/\sqrt{2}, 1/\sqrt{2}),\ 1.443\,I)$. The 20 values of m are evenly spaced in $[1.2, 2.85]$. The sources are widened along opposite axes and shifted in opposite directions, while P^w is isotropic, broader than the base, and ends shifted perpendicular to the offsets of the sources.

Gaussian experiments. We validate the diagnostic in a parametric setting where both source-estimation error and composition error can be computed directly. For each configuration, each source is refitted by maximum likelihood as a full-covariance Gaussian from $n = 20{,}000$ fresh samples, using the sample covariance with divisor n. For Gaussian precisions, the fitted composition is obtained in closed form from (11). We use $R = 2000$ independent refits per configuration.

Error amplification for replicate r is

$$A_r = \frac{\mathrm{TV}(P^w, \widehat{P}^w_r)}{\sum_{i\in\{0,a_1,a_2\}} \mathrm{TV}(P^i, \widehat{P}^i_r)},$$

and we report the mean over replicates. If the composed precision $\widehat{\Lambda}^w$ is not positive definite, the fitted composition is not a valid distribution and that replicate is omitted. This occurs only in the

$r > 2$ anisotropic family. Each TV distance is estimated on 8192 samples from the corresponding true distribution, drawn once per configuration and held fixed across refits, using

$$\mathrm{TV}(P, \widehat{P}) = \mathbb{E}_{X\sim P}\left[\max\left(0, 1 - \frac{\widehat{p}(X)}{p(X)}\right)\right].$$

Additionally, we note that the sensitivity bound involving κ measures source error using $\|u_i\|_{L^2(P^i)}$ rather than TV, so $\widehat{\mathrm{rESS}}^{-1/2}$ is not strictly an upper bound on this notion of amplification A. However, across all configurations, $\sum_i \|u_i\|_{L^2(P^i)} / \sum_i \mathrm{TV}(P^i, \widehat{P}^i)$ lies between 2.81 and 2.88, so the two quantities differ by a nearly constant factor, so the same intuition holds for Figure 4.

The diagnostic is evaluated separately using exact Gaussian densities on $N = 20{,}000$ fresh samples per source, with 12 independent draws per configuration. The bound figure plots $\widehat{\mathrm{rESS}}^{-1/2}$ against A, using the median $-\log \widehat{\mathrm{rESS}}$ over the 12 draws for each configuration. For each Gaussian family, we also compute the Spearman correlation between $\widehat{\mathrm{rESS}}^{-1/2}$ and A across its 20 configurations separately for each draw, and display the median and full range of these 12 correlations (Figure 4).

Image density-ratio estimation. For the image experiments, the required density ratios are estimated using the latent diffusion models described in App. I.5. We take this approach because we observe density ratios to be uninformative in pixel space, since density ratio estimation suffers from the curse of dimensionality and is statistically challenging in high dimensions.

As discussed in App. G.8, the image models do not provide $\log P^{a_j}(x) - \log P^0(x)$ in closed form. We estimate these ratios using differences of the condition-dependent DDPM variational bounds,

$$\log P^a(x) - \log P^0(x) \approx \mathcal{L}_0(x) - \mathcal{L}_a(x),$$

where

$$\mathcal{L}_a(x) = \sum_{j\in\mathcal{T}} \frac{\beta_j}{2(1-\beta_j)(1-\bar{\alpha}_{j-1})}\, \mathbb{E}_\epsilon \left\|\epsilon - \epsilon_a(x_j, j)\right\|^2,$$

and

$$x_j = \sqrt{\bar{\alpha}_j}\, x + \sqrt{1-\bar{\alpha}_j}\, \epsilon.$$

This is the finite-step variational DDPM bound of Ho et al. (2020) for timesteps $\mathcal{T}$. Because the models predict v, we recover the noise prediction as $\epsilon_a = \alpha_j v_a + \sigma_j x_j$. Following Li et al. (2023), the same (j, ϵ, x_j) is used for both conditions, and their integrands are subtracted before averaging.

We evaluate both a trimmed window $\mathcal{T} = \{100, \ldots, 900\}$ and an unconstrained window $\mathcal{T} = \{1, \ldots, 997\}$ of the 1000-step schedule. We exclude timestep 0, which contains the reconstruction term, and timesteps 998 and 999, for which the terminal zero signal-to-noise ratio produces an unbounded log-SNR interval.

For both rooms and CelebA, we use $N = 1024$ held-out samples per source and $M = 512$ Monte Carlo draws per sample under a fixed seed. We sample uniformly in log-SNR over the intervals corresponding to $\mathcal{T}$ and weight by the inverse interval width so that the estimator is unbiased for the discrete sum. The draws use randomly shifted van der Corput points.

J Additional room images

In this section, we provide additional samples from the compositions of learned diffusion models. These images are shown in Figure 5.

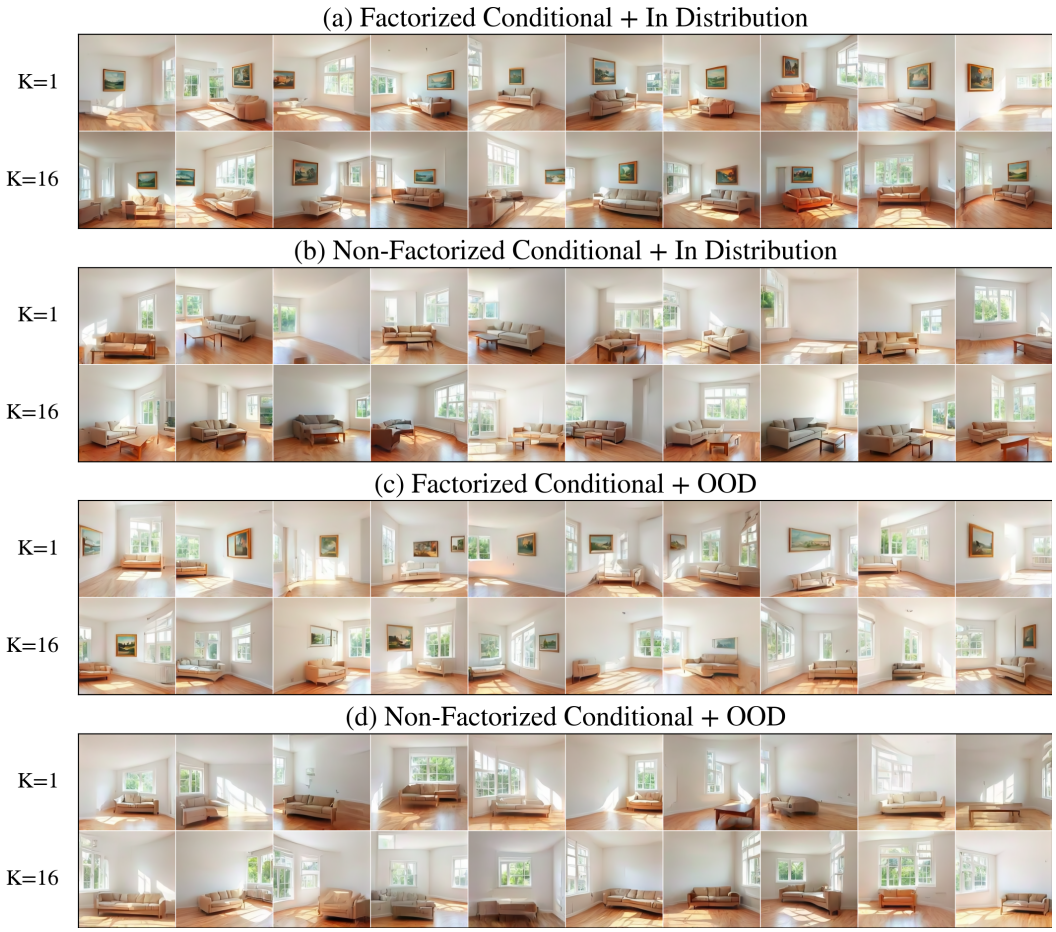


Figure 5: Additional samples from composed conditional diffusion models. Setup and trends are the same as reported in Figure 2.

K Tables

K.1 Two-dimensional Gaussian (conditional model)

Table 18: **Factorized Conditionals + In-Distribution Composition.** $P^{a_1} = \mathcal{N}(0,\, \mathrm{diag}(10, 1))$, $P^{a_2} = \mathcal{N}(0,\, \mathrm{diag}(1, 10))$, $P^0 = \mathcal{N}(0,\, 10I)$, $\mathbf{P}^w = \mathcal{N}(0,\, I)$): SW_2 (mean $\pm$ std over 30 runs)

$K \backslash N$	N=100	N=1000	N=10000	analytical
K=1	0.3537 ± 0.0859	0.1089 ± 0.0266	0.0747 ± 0.0231	0.0354 ± 0.0056
K=4	0.4276 ± 0.1379	0.0974 ± 0.0308	0.0750 ± 0.0208	0.0376 ± 0.0054
K=16	0.4894 ± 0.1747	0.0956 ± 0.0271	0.0772 ± 0.0226	0.0351 ± 0.0054
K=64	0.5136 ± 0.1882	0.0937 ± 0.0269	0.0758 ± 0.0223	0.0356 ± 0.0044
K=256	0.5261 ± 0.1913	0.1010 ± 0.0274	0.0777 ± 0.0237	0.0346 ± 0.0057

Table 19: **Factorized Conditionals + In-Distribution Composition.** $P^{a_1} = \mathcal{N}(0,\, \mathrm{diag}(10, 1))$, $P^{a_2} = \mathcal{N}(0,\, \mathrm{diag}(1, 10))$, $P^0 = \mathcal{N}(0,\, 10I)$, $\mathbf{P}^w = \mathcal{N}(0,\, I)$): MMD^2 (mean $\pm$ std over 30 runs)

$K \backslash N$	N=100	N=1000	N=10000	analytical
K=1	0.0252 ± 0.0173	0.0025 ± 0.0014	0.0011 ± 0.0009	-0.0000 ± 0.0001
K=4	0.0449 ± 0.0397	0.0021 ± 0.0015	0.0010 ± 0.0008	0.0000 ± 0.0001
K=16	0.0604 ± 0.0526	0.0019 ± 0.0014	0.0012 ± 0.0009	-0.0000 ± 0.0001
K=64	0.0657 ± 0.0561	0.0018 ± 0.0015	0.0012 ± 0.0010	-0.0000 ± 0.0001
K=256	0.0684 ± 0.0564	0.0021 ± 0.0015	0.0012 ± 0.0011	-0.0000 ± 0.0001

Table 20: **Non-Factorized Conditionals + In-Distribution Composition.** $P^{a_1} = \mathcal{N}(0,\, \mathrm{diag}(10, 1))$, $P^{a_2} = \mathcal{N}(0,\, \mathrm{diag}(1, 10))$, $P^0 = \mathcal{N}(0,\, 20I)$, $\mathbf{P}^w = \mathcal{N}(0,\, \frac{20}{21}I)$): SW_2 (mean $\pm$ std over 30 runs)

$K \backslash N$	N=100	N=1000	N=10000	analytical
K=1	0.4454 ± 0.1490	0.1234 ± 0.0310	0.0957 ± 0.0240	0.0655 ± 0.0085
K=4	0.5331 ± 0.1920	0.0969 ± 0.0293	0.0751 ± 0.0219	0.0412 ± 0.0045
K=16	0.5914 ± 0.2111	0.0901 ± 0.0253	0.0750 ± 0.0227	0.0353 ± 0.0061
K=64	0.6111 ± 0.2162	0.0905 ± 0.0237	0.0740 ± 0.0230	0.0357 ± 0.0051
K=256	0.6216 ± 0.2201	0.0963 ± 0.0270	0.0793 ± 0.0234	0.0333 ± 0.0043

Table 21: **Non-Factorized Conditionals + In-Distribution Composition.** $P^{a_1} = \mathcal{N}(0,\, \mathrm{diag}(10, 1))$, $P^{a_2} = \mathcal{N}(0,\, \mathrm{diag}(1, 10))$, $P^0 = \mathcal{N}(0,\, 20I)$, $\mathbf{P}^w = \mathcal{N}(0,\, \frac{20}{21}I)$): MMD^2 (mean $\pm$ std over 30 runs)

$K \backslash N$	N=100	N=1000	N=10000	analytical
K=1	0.0507 ± 0.0412	0.0038 ± 0.0021	0.0020 ± 0.0011	0.0007 ± 0.0003
K=4	0.0803 ± 0.0628	0.0022 ± 0.0016	0.0011 ± 0.0009	0.0001 ± 0.0001
K=16	0.0993 ± 0.0713	0.0018 ± 0.0012	0.0012 ± 0.0010	0.0000 ± 0.0001
K=64	0.1057 ± 0.0748	0.0018 ± 0.0013	0.0012 ± 0.0011	0.0000 ± 0.0001
K=256	0.1093 ± 0.0764	0.0021 ± 0.0016	0.0013 ± 0.0012	-0.0000 ± 0.0001

Table 22: **Factorized Conditionals + Out-Of-Distribution Composition.** $P^{a_1} = \mathcal{N}(0,\, \mathrm{diag}(10, 1))$, $P^{a_2} = \mathcal{N}(0,\, \mathrm{diag}(1, 10))$, $P^0 = \mathcal{N}(0,\, I)$, $\mathbf{P}^w = \mathcal{N}(0,\, 10I)$): SW_2 (mean $\pm$ std over 30 runs)

$K \backslash N$	N=100	N=1000	N=10000	analytical
K=1	0.9658 ± 0.2036	0.5577 ± 0.1655	0.3948 ± 0.1459	0.1122 ± 0.0189
K=4	2.0688 ± 0.6346	1.6885 ± 0.4857	1.1773 ± 0.4160	0.1130 ± 0.0129
K=16	2.9177 ± 0.9132	2.7214 ± 0.8497	1.9699 ± 0.6611	0.1133 ± 0.0228
K=64	3.3210 ± 1.0273	3.3348 ± 1.0738	2.4983 ± 0.8347	0.1124 ± 0.0147
K=256	3.5097 ± 1.0678	3.6426 ± 1.2007	2.8210 ± 0.9801	0.1140 ± 0.0158

Table 23: **Factorized Conditionals + Out-Of-Distribution Composition.** $P^{a_1} = \mathcal{N}(0,\, \mathrm{diag}(10, 1))$, $P^{a_2} = \mathcal{N}(0,\, \mathrm{diag}(1, 10))$, $P^0 = \mathcal{N}(0,\, I)$, $\mathbf{P}^w = \mathcal{N}(0,\, 10I)$): MMD^2 (mean $\pm$ std over 30 runs)

$K \backslash N$	N=100	N=1000	N=10000	analytical
K=1	0.0135 ± 0.0067	0.0042 ± 0.0025	0.0025 ± 0.0019	-0.0000 ± 0.0001
K=4	0.0872 ± 0.0513	0.0541 ± 0.0264	0.0271 ± 0.0160	-0.0000 ± 0.0001
K=16	0.1677 ± 0.0977	0.1294 ± 0.0694	0.0695 ± 0.0373	0.0000 ± 0.0002
K=64	0.2039 ± 0.1171	0.1778 ± 0.0930	0.1042 ± 0.0538	-0.0000 ± 0.0001
K=256	0.2210 ± 0.1253	0.2013 ± 0.1010	0.1269 ± 0.0685	0.0000 ± 0.0001

Table 24: **Non-Factorized Conditionals + Out-Of-Distribution Composition.** $P^{a_1} = \mathcal{N}(0,\, \mathrm{diag}(10, 1))$, $P^{a_2} = \mathcal{N}(0,\, \mathrm{diag}(1, 10))$, $P^0 = \mathcal{N}(0,\, 1.1I)$, $\mathbf{P}^w = \mathcal{N}(0,\, \frac{110}{21}I)$): SW_2 (mean $\pm$ std over 30 runs)

$K \backslash N$	N=100	N=1000	N=10000	analytical
K=1	1.0957 ± 0.3179	0.9154 ± 0.2265	0.8518 ± 0.2079	0.6375 ± 0.0257
K=4	2.1611 ± 0.6579	1.8306 ± 0.5788	1.3763 ± 0.5356	0.1764 ± 0.0222
K=16	3.0011 ± 0.8826	2.8830 ± 0.9744	2.0117 ± 0.8550	0.0981 ± 0.0197
K=64	3.4224 ± 1.0159	3.6247 ± 1.2177	2.5011 ± 1.1307	0.0831 ± 0.0128
K=256	3.6076 ± 1.0687	3.9971 ± 1.3265	2.8322 ± 1.3633	0.0820 ± 0.0147

Table 25: **Non-Factorized Conditionals + Out-Of-Distribution Composition.** $P^{a_1} = \mathcal{N}(0,\, \mathrm{diag}(10, 1))$, $P^{a_2} = \mathcal{N}(0,\, \mathrm{diag}(1, 10))$, $P^0 = \mathcal{N}(0,\, 1.1I)$, $\mathbf{P}^w = \mathcal{N}(0,\, \frac{110}{21}I)$): MMD^2 (mean $\pm$ std over 30 runs)

$K \backslash N$	N=100	N=1000	N=10000	analytical
K=1	0.0300 ± 0.0148	0.0189 ± 0.0079	0.0172 ± 0.0063	0.0117 ± 0.0010
K=4	0.1267 ± 0.0599	0.0770 ± 0.0358	0.0438 ± 0.0271	0.0009 ± 0.0003
K=16	0.2241 ± 0.1033	0.1758 ± 0.0890	0.0862 ± 0.0539	0.0001 ± 0.0002
K=64	0.2689 ± 0.1243	0.2498 ± 0.1201	0.1215 ± 0.0794	-0.0000 ± 0.0001
K=256	0.2866 ± 0.1310	0.2853 ± 0.1313	0.1455 ± 0.0996	-0.0000 ± 0.0001

K.2 Two-dimensional Gaussian (separate models)

Table 26: **Factorized Conditionals + In-Distribution Composition.** Separate models, $P^{a_1} = \mathcal{N}(0,\, \mathrm{diag}(10, 1))$, $P^{a_2} = \mathcal{N}(0,\, \mathrm{diag}(1, 10))$, $P^0 = \mathcal{N}(0,\, 10I)$, $\mathbf{P}^w = \mathcal{N}(0,\, I)$): SW_2 (mean $\pm$ std over 30 runs)

$K \backslash N$	N=100	N=1000	N=10000	analytical
K=1	0.4443 ± 0.0893	0.1339 ± 0.0375	0.0933 ± 0.0327	0.0370 ± 0.0051
K=4	0.5546 ± 0.1553	0.1217 ± 0.0304	0.0765 ± 0.0259	0.0363 ± 0.0057
K=16	0.6076 ± 0.1804	0.1177 ± 0.0284	0.0763 ± 0.0224	0.0372 ± 0.0067
K=64	0.6267 ± 0.1911	0.1140 ± 0.0327	0.0743 ± 0.0278	0.0365 ± 0.0048
K=256	0.6364 ± 0.1947	0.1196 ± 0.0291	0.0786 ± 0.0242	0.0359 ± 0.0062

Table 27: **Factorized Conditionals + In-Distribution Composition.** Separate models, $P^{a_1} = \mathcal{N}(0,\, \mathrm{diag}(10, 1))$, $P^{a_2} = \mathcal{N}(0,\, \mathrm{diag}(1, 10))$, $P^0 = \mathcal{N}(0,\, 10I)$, $\mathbf{P}^w = \mathcal{N}(0,\, I)$): MMD^2 (mean $\pm$ std over 30 runs)

$K \backslash N$	N=100	N=1000	N=10000	analytical
K=1	0.0323 ± 0.0206	0.0041 ± 0.0026	0.0019 ± 0.0016	0.0000 ± 0.0001
K=4	0.0647 ± 0.0485	0.0033 ± 0.0021	0.0012 ± 0.0012	0.0000 ± 0.0001
K=16	0.0817 ± 0.0584	0.0030 ± 0.0017	0.0011 ± 0.0008	0.0000 ± 0.0001
K=64	0.0876 ± 0.0624	0.0028 ± 0.0019	0.0011 ± 0.0011	-0.0000 ± 0.0001
K=256	0.0903 ± 0.0638	0.0031 ± 0.0019	0.0012 ± 0.0011	0.0000 ± 0.0001

Table 28: **Non-Factorized Conditionals + In-Distribution Composition.** Separate models, $P^{a_1} = \mathcal{N}(0,\, \mathrm{diag}(10, 1))$, $P^{a_2} = \mathcal{N}(0,\, \mathrm{diag}(1, 10))$, $P^0 = \mathcal{N}(0,\, 20I)$, $\mathbf{P}^w = \mathcal{N}(0,\, \frac{20}{21}I)$): SW_2 (mean $\pm$ std over 30 runs)

$K \backslash N$	N=100	N=1000	N=10000	analytical
K=1	0.5553 ± 0.1378	0.1437 ± 0.0343	0.1106 ± 0.0300	0.0673 ± 0.0073
K=4	0.6524 ± 0.1708	0.1213 ± 0.0313	0.0781 ± 0.0255	0.0389 ± 0.0079
K=16	0.6930 ± 0.1783	0.1148 ± 0.0283	0.0711 ± 0.0194	0.0356 ± 0.0057
K=64	0.7034 ± 0.1815	0.1147 ± 0.0322	0.0705 ± 0.0248	0.0358 ± 0.0049
K=256	0.7110 ± 0.1810	0.1157 ± 0.0320	0.0737 ± 0.0226	0.0350 ± 0.0063

Table 29: **Non-Factorized Conditionals + In-Distribution Composition.** Separate models, $P^{a_1} = \mathcal{N}(0,\, \mathrm{diag}(10, 1))$, $P^{a_2} = \mathcal{N}(0,\, \mathrm{diag}(1, 10))$, $P^0 = \mathcal{N}(0,\, 20I)$, $\mathbf{P}^w = \mathcal{N}(0,\, \frac{20}{21}I)$): MMD^2 (mean $\pm$ std over 30 runs)

$K \backslash N$	N=100	N=1000	N=10000	analytical
K=1	0.0640 ± 0.0471	0.0053 ± 0.0029	0.0030 ± 0.0018	0.0007 ± 0.0002
K=4	0.0979 ± 0.0708	0.0037 ± 0.0023	0.0014 ± 0.0012	0.0001 ± 0.0002
K=16	0.1123 ± 0.0763	0.0031 ± 0.0020	0.0010 ± 0.0007	0.0000 ± 0.0001
K=64	0.1159 ± 0.0789	0.0032 ± 0.0022	0.0010 ± 0.0010	-0.0000 ± 0.0001
K=256	0.1185 ± 0.0796	0.0032 ± 0.0024	0.0011 ± 0.0010	0.0000 ± 0.0001

Table 30: **Factorized Conditionals + Out-Of-Distribution Composition.** Separate models, $P^{a_1} = \mathcal{N}(0,\, \mathrm{diag}(10, 1))$, $P^{a_2} = \mathcal{N}(0,\, \mathrm{diag}(1, 10))$, $P^0 = \mathcal{N}(0,\, I)$, $\mathbf{P}^w = \mathcal{N}(0,\, 10I)$): SW_2 (mean $\pm$ std over 30 runs)

$K \backslash N$	N=100	N=1000	N=10000	analytical
K=1	1.1771 ± 0.3766	0.8397 ± 0.2398	0.7643 ± 0.3888	0.1163 ± 0.0150
K=4	2.2327 ± 0.9720	2.0231 ± 0.6790	2.0465 ± 0.9039	0.1160 ± 0.0205
K=16	3.1107 ± 1.4567	3.2394 ± 0.9808	3.4382 ± 1.6403	0.1174 ± 0.0206
K=64	3.5605 ± 1.6860	3.9171 ± 1.0955	4.2961 ± 2.1283	0.1143 ± 0.0150
K=256	3.8044 ± 1.8279	4.2434 ± 1.2064	4.7380 ± 2.4489	0.1061 ± 0.0211

Table 31: **Factorized Conditionals + Out-Of-Distribution Composition.** Separate models, $P^{a_1} = \mathcal{N}(0,\, \mathrm{diag}(10, 1))$, $P^{a_2} = \mathcal{N}(0,\, \mathrm{diag}(1, 10))$, $P^0 = \mathcal{N}(0,\, I)$, $\mathbf{P}^w = \mathcal{N}(0,\, 10I)$): MMD^2 (mean $\pm$ std over 30 runs)

$K \backslash N$	N=100	N=1000	N=10000	analytical
K=1	0.0179 ± 0.0090	0.0083 ± 0.0042	0.0049 ± 0.0029	0.0000 ± 0.0001
K=4	0.0903 ± 0.0639	0.0678 ± 0.0387	0.0655 ± 0.0354	0.0000 ± 0.0001
K=16	0.1727 ± 0.1192	0.1600 ± 0.0751	0.1777 ± 0.1020	0.0000 ± 0.0001
K=64	0.2171 ± 0.1387	0.2174 ± 0.0922	0.2428 ± 0.1353	-0.0000 ± 0.0001
K=256	0.2403 ± 0.1440	0.2445 ± 0.1040	0.2707 ± 0.1440	-0.0000 ± 0.0001

Table 32: **Non-Factorized Conditionals + Out-Of-Distribution Composition.** Separate models, $P^{a_1} = \mathcal{N}(0,\, \mathrm{diag}(10, 1))$, $P^{a_2} = \mathcal{N}(0,\, \mathrm{diag}(1, 10))$, $P^0 = \mathcal{N}(0,\, 1.1I)$, $\mathbf{P}^w = \mathcal{N}(0,\, \frac{110}{21}I)$): SW_2 (mean $\pm$ std over 30 runs)

$K\backslash N$	N=100	N=1000	N=10000	analytical
K=1	1.2463 ± 0.4880	1.0431 ± 0.3266	0.8929 ± 0.4140	0.6418 ± 0.0308
K=4	2.3105 ± 0.9665	2.2133 ± 0.8139	1.8170 ± 0.9345	0.1704 ± 0.0218
K=16	3.1794 ± 1.4174	3.4853 ± 1.2953	2.9968 ± 1.5100	0.0905 ± 0.0153
K=64	3.6653 ± 1.6599	4.2021 ± 1.5352	3.8652 ± 1.8677	0.0821 ± 0.0100
K=256	3.8937 ± 1.7769	4.5325 ± 1.6601	4.4165 ± 2.1453	0.0815 ± 0.0139

Table 33: **Non-Factorized Conditionals + Out-Of-Distribution Composition.** Separate models, $P^{a_1} = \mathcal{N}(0,\, \mathrm{diag}(10, 1))$, $P^{a_2} = \mathcal{N}(0,\, \mathrm{diag}(1, 10))$, $P^0 = \mathcal{N}(0,\, 1.1I)$, $\mathbf{P}^w = \mathcal{N}(0,\, \frac{110}{21}I)$): MMD^2 (mean $\pm$ std over 30 runs)

$K\backslash N$	N=100	N=1000	N=10000	analytical
K=1	0.0340 ± 0.0180	0.0202 ± 0.0090	0.0156 ± 0.0090	0.0119 ± 0.0012
K=4	0.1261 ± 0.0669	0.1034 ± 0.0566	0.0736 ± 0.0543	0.0008 ± 0.0003
K=16	0.2251 ± 0.1118	0.2212 ± 0.1242	0.1741 ± 0.1179	0.0001 ± 0.0001
K=64	0.2814 ± 0.1346	0.2811 ± 0.1461	0.2483 ± 0.1399	-0.0000 ± 0.0001
K=256	0.3058 ± 0.1422	0.3061 ± 0.1527	0.2891 ± 0.1472	-0.0000 ± 0.0001

K.3 Gaussian mixture models (conditional model)

Table 34: **GMM, In-Distribution** (SW_2 mean $\pm$ std over 30 runs)

$K\backslash N$	N=100	N=1000	N=10000	analytical
K=1	1.2113 ± 0.2110	1.2470 ± 0.1079	1.1704 ± 0.0914	1.2133 ± 0.0310
K=4	0.6678 ± 0.2758	0.2842 ± 0.0728	0.2318 ± 0.0606	0.2414 ± 0.0308
K=16	0.7735 ± 0.3478	0.2206 ± 0.0775	0.1454 ± 0.0567	0.1044 ± 0.0314
K=64	0.8130 ± 0.3713	0.2134 ± 0.0730	0.1478 ± 0.0533	0.0736 ± 0.0266
K=256	0.8288 ± 0.3826	0.1985 ± 0.0697	0.1350 ± 0.0539	0.0771 ± 0.0285

Table 35: **GMM, In-Distribution** (MMD^2 mean $\pm$ std over 30 runs)

$K\backslash N$	N=100	N=1000	N=10000	analytical
K=1	0.0590 ± 0.0272	0.0457 ± 0.0097	0.0370 ± 0.0088	0.0392 ± 0.0026
K=4	0.0576 ± 0.0474	0.0039 ± 0.0022	0.0016 ± 0.0019	0.0002 ± 0.0002
K=16	0.0839 ± 0.0702	0.0045 ± 0.0031	0.0015 ± 0.0020	0.0000 ± 0.0002
K=64	0.0935 ± 0.0791	0.0045 ± 0.0032	0.0016 ± 0.0023	-0.0000 ± 0.0001
K=256	0.0977 ± 0.0825	0.0040 ± 0.0028	0.0017 ± 0.0021	-0.0000 ± 0.0002

Table 36: **GMM, Out-of-Distribution** (SW_2 mean $\pm$ std over 30 runs)

$K\backslash N$	N=100	N=1000	N=10000	analytical
K=1	4.0336 ± 0.5817	4.0508 ± 0.3100	3.9606 ± 0.2339	4.2102 ± 0.0441
K=4	2.7357 ± 0.6423	2.7265 ± 0.4211	2.4675 ± 0.3453	2.5081 ± 0.0456
K=16	2.4276 ± 0.7330	2.4556 ± 0.6946	2.1718 ± 0.4914	1.7704 ± 0.0374
K=64	2.6212 ± 0.8565	2.7205 ± 0.9299	2.4005 ± 0.6674	1.5531 ± 0.0455
K=256	3.0238 ± 1.0936	3.0562 ± 1.1639	2.6706 ± 0.8479	1.4670 ± 0.0506

Table 37: **GMM, Out-of-Distribution** (MMD^2 mean $\pm$ std over 30 runs)

$K\backslash N$	N=100	N=1000	N=10000	analytical
K=1	0.2616 ± 0.0777	0.2343 ± 0.0382	0.2262 ± 0.0347	0.2692 ± 0.0077
K=4	0.2208 ± 0.1082	0.0892 ± 0.0303	0.0583 ± 0.0202	0.0362 ± 0.0027
K=16	0.2514 ± 0.1312	0.1009 ± 0.0450	0.0610 ± 0.0322	0.0102 ± 0.0009
K=64	0.2782 ± 0.1413	0.1539 ± 0.0785	0.0879 ± 0.0554	0.0067 ± 0.0008
K=256	0.3068 ± 0.1460	0.2036 ± 0.1022	0.1133 ± 0.0767	0.0056 ± 0.0008

K.4 Gaussian mixture models (separate models)

Table 38: **GMM, In-Distribution** Separate models, (SW_2 mean $\pm$ std over 30 runs)

$K\backslash N$	N=100	N=1000	N=10000	analytical
K=1	1.2393 ± 0.2418	1.2108 ± 0.1863	1.2029 ± 0.1275	1.2157 ± 0.0305
K=4	0.7108 ± 0.2359	0.4015 ± 0.1510	0.2925 ± 0.1031	0.2452 ± 0.0317
K=16	0.8054 ± 0.3097	0.4442 ± 0.2769	0.2116 ± 0.1222	0.1128 ± 0.0311
K=64	0.8440 ± 0.3406	0.5025 ± 0.4534	0.2196 ± 0.1230	0.0843 ± 0.0251
K=256	0.8624 ± 0.3564	0.5514 ± 0.6366	0.2214 ± 0.1321	0.0740 ± 0.0336

Table 39: **GMM, In-Distribution** Separate models, (MMD^2 mean $\pm$ std over 30 runs)

$K\backslash N$	N=100	N=1000	N=10000	analytical
K=1	0.0630 ± 0.0308	0.0444 ± 0.0176	0.0419 ± 0.0119	0.0396 ± 0.0027
K=4	0.0622 ± 0.0525	0.0104 ± 0.0110	0.0053 ± 0.0059	0.0002 ± 0.0002
K=16	0.0864 ± 0.0743	0.0183 ± 0.0238	0.0056 ± 0.0099	0.0000 ± 0.0002
K=64	0.0969 ± 0.0862	0.0233 ± 0.0323	0.0064 ± 0.0110	0.0000 ± 0.0002
K=256	0.1014 ± 0.0921	0.0269 ± 0.0383	0.0066 ± 0.0127	-0.0000 ± 0.0001

Table 40: **GMM, Out-of-Distribution** Separate models, (SW_2 mean $\pm$ std over 30 runs)

$K\backslash N$	N=100	N=1000	N=10000	analytical
K=1	4.2381 ± 0.6608	3.8951 ± 0.3696	3.8501 ± 0.2915	4.2093 ± 0.0472
K=4	2.7454 ± 0.7513	2.3919 ± 0.4998	2.2963 ± 0.2614	2.5141 ± 0.0447
K=16	2.2444 ± 0.8758	1.8197 ± 0.6064	1.6168 ± 0.3447	1.7899 ± 0.0377
K=64	2.5258 ± 2.1074	1.6895 ± 0.6533	1.4262 ± 0.3536	1.5467 ± 0.0364
K=256	2.7293 ± 3.1365	1.6850 ± 0.7136	1.3582 ± 0.3864	1.4619 ± 0.0493

Table 41: **GMM, Out-of-Distribution** Separate models, (MMD^2 mean $\pm$ std over 30 runs)

$K\backslash N$	N=100	N=1000	N=10000	analytical
K=1	0.2727 ± 0.0964	0.2484 ± 0.0501	0.2482 ± 0.0373	0.2706 ± 0.0076
K=4	0.2555 ± 0.1423	0.1006 ± 0.0552	0.0559 ± 0.0219	0.0365 ± 0.0025
K=16	0.3063 ± 0.1723	0.1165 ± 0.0949	0.0368 ± 0.0221	0.0106 ± 0.0010
K=64	0.3283 ± 0.1820	0.1378 ± 0.1121	0.0368 ± 0.0238	0.0066 ± 0.0007
K=256	0.3383 ± 0.1888	0.1585 ± 0.1274	0.0407 ± 0.0273	0.0056 ± 0.0008