UNDERSTANDING MODE COLLAPSE IN DIFFUSION MODELS UNDER CLASSIFIER-FREE GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

The use of classifier-free guidance in diffusion and flow models has proven critical to their empirical success, as it allows for the controllable generation of lower-temperature and higher-fidelity samples. Despite its widespread use, classifier-free guidance remains both empirically and theoretically poorly understood, owing to the difficulty of solving the associated transport equations for arbitrary normalizing flows. In this paper, we observe that the effects of classifier-free guidance arise from a mismatch between the representation capabilities of the conditional and unconditional distributions, wherein the conditional distribution exhibits higher complexity and fidelity. We propose an analytical model to capture this observation: the conditional distribution is a Gaussian mixture whose means lie on an ellipsoid, and the unconditional distribution is a single Gaussian approximating the mixture. The flows arising from this mixture are closed under scaling, enabling us to derive analytic closed-form solutions for the guided distribution at all noise scales and guidance levels. Our analytic solution shows that classifier-free guidance induces (a) a shift along a mean-axis direction and (b) a persistent skew toward, and concentration around, a singular point corresponding to the conditional sample starting from zero noise. We validate our theoretical predictions on pretrained, large-scale image diffusion models.

1 INTRODUCTION

Over the last several years, diffusion models have become the dominant paradigm for generative modeling across a wide spectrum of continuous data modalities, ranging from image (Zheng et al., 2026), audio, and video generation (Yang et al., 2025) tasks to computational biology (Abramson et al., 2024), robotics (Wang et al., 2022), and time-series forecasting problems (Ding et al., 2026). The fundamental principle that unifies diffusion-style generation methods is the construction of a normalizing flow that smoothly interpolates between data samples and random noise. The resulting flow can easily be learned by minimizing a standard supervised loss and sampled using either stochastic (SDE-based) (Ho et al., 2020) or deterministic (ODE-based) (Song et al., 2020a) sampling methodologies.

A crucial component in the success of diffusion across these domains is the ability to amplify the performance of a conditional generative model via the construction of *guided* flows (Ho & Salimans, 2022), which linearly combine the learned flow of the unconditional and conditional distributions at inference time to enhance the features of the conditional distribution. The adoption of guidance extends even beyond diffusion models, as various other generative paradigms have reported performance improvements by including guidance-style contrastive terms in their formulation (Sanchez et al., 2023; Chang et al., 2023; Nisonoff et al., 2025).

While guided flows generally produce both lower-temperature and higher-quality samples than either their unconditional or conditional counterparts, they break the fundamental data-to-noise interpolative foundation on which the theory of diffusion models is built (Du et al., 2023). As a result, the effects of guidance on the resulting sample distribution remain comparatively poorly understood. Previous work has identified the ability of guidance to selectively amplify or suppress specific feature directions (Li et al., 2026) and suggests guided flows weakly approximate exponentially-weighted density products of the underlying flows (Liu et al., 2022; Du et al., 2023; Bradley & Nakkiran, 2024). The derivation of exact closed-form solutions has previously necessitated either (1) restricting the

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

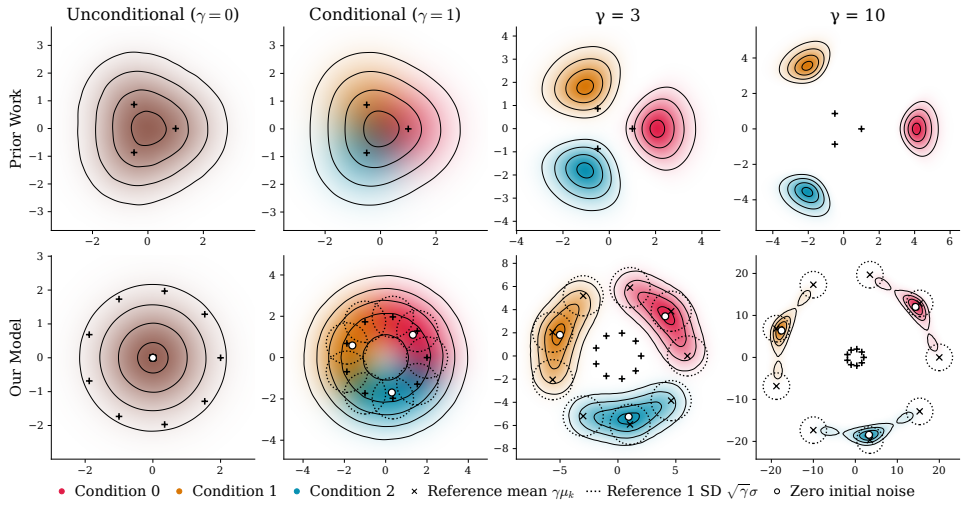

Figure 1: **Top:** A standard model for classifier-free guidance (Ho & Salimans, 2022), where the conditional distribution is a single component of the full unconditional mixture distribution. **Bottom:** Our model, where each conditional distribution is a Gaussian mixture (pictured here with three components per condition) and the unconditional is a single Gaussian approximating the marginalized mixture. Our model admits an analytical closed-form solution, capturing the *feature-magnification* and *modality collapse* effects outlined in Theorem 3.4.

analysis to Gaussian distributions, for which the underlying flows are purely linear (Wu et al., 2024; Li et al., 2026), or (2) asymptotic analysis or integration of a correction term over the associated sample trajectories (Zheng & Lan, 2023; Skreta et al., 2025; Jiang & Ma, 2026; Pavasovic et al., 2025). For an extended discussion of the prior work, see Appendix A.

Contributions. In this work, we make several key advancements in the understanding of classifier free guidance from both theoretical and empirical perspectives:

1. We derive what, to our knowledge, is the first closed-form solution for a guided density involving a multi-modal conditional distribution, for which the underlying flow is nonlinear.
2. We consider a model for guidance where the conditional distribution is a mixture, while the unconditional distribution is represented by a single Gaussian. We argue that this better represents classifier-free guidance, as in practice unconditionally-generated samples exhibit poor fidelity, suggesting that unconditional flows are less expressive. Our setting also captures techniques such as self-guidance (Li et al., 2025) and auto-guidance (Karras et al., 2024), where a lower-capacity conditional model is used in place of the unconditional flow.
3. Our analysis exactly predicts both (1) the simultaneous magnification of features across different, component-specific axes and (2) concentration around a single point represented by setting the initial noise to zero. We additionally show that this mode-collapse effect is *bounded* and derive a limit on the concentration around the collapsed mode.

We empirically validate our findings on pretrained, large-scale image diffusion models for text- and class-conditioned settings. We show the prevalence of both the hypothesized cluster-specific feature magnification effect, in addition to a limited degree of modality collapse to the mode represented by the zero-initialized sample, in the specific manners predicted by our model.

2 Diffusion Preliminaries

We base our formulation of diffusion on the stochastic interpolant framework of Albergo & Vanden-Eijnden (2022); Albergo et al. (2025). In this framework, we construct a random variable X_t that interpolates between samples from a data distribution $X_0 \sim p(x)$ and pure noise X_1. We consider the standard linear interpolation given by a *diffusion schedule* $\alpha(t), \sigma(t)$ weighting the data and noise respectively, for which the interpolation is equivalent to the standard diffusion forward process used in Ho et al. (2020); Lipman et al. (2022); Zheng et al. (2023); Song et al. (2020a) and related literature.

Notation: For $A \succ 0$, we use $\|\cdot\|_A$ to denote the Mahalanobis norm $\|z\|_A := \sqrt{z^\top A z}$. We use uppercase ($X$) to denote matrices or random variables and lowercase (x) for vectors or scalars.

Definition 2.1 (Noising Process)**.** Let X_0 be a continuous random variable supported on $\mathbb{R}^n$ with density $p_0(x)$ and finite second moment, and let $X_1 \sim \mathcal{N}(0, \Sigma)$ be independent of X_0, where $\Sigma \in \mathbb{R}^{n\times n}$ is symmetric positive definite. For continuously differentiable functions $\alpha, \sigma\colon [0,1] \to \mathbb{R}$, define,

$$X_t := \alpha(t)X_0 + \sigma(t)X_1, \tag{2.1}$$

where the noise schedule (α, σ) satisfies,

$$\alpha(0) = 1, \quad \sigma(0) = 0, \qquad \alpha(1) = 0, \quad \sigma(1) = 1, \qquad \alpha(t)^2 + \sigma(t)^2 = 1, \tag{2.2}$$

with α non-increasing and σ non-decreasing. For simplicity we only consider variance-preserving interpolants such that $\alpha(t)^2 + \sigma(t)^2 = 1$ for all t.

The terminal distribution X_1 is generally taken to be $\mathcal{N}(0, I)$, i.e. $\Sigma = I$. We leave Σ generic in our formulation as the added flexibility simplifies the proof of our main result, Theorem 3.4, which relies on the construction of a flow with a misspecified $p_1(x)$ distribution.

Proposition 2.2 (Transport equation, Albergo et al., 2025)**.** *The densities $p_t(x)$ associated with X_t as defined in Definition 2.1 satisfy the continuity equation,*

$$\partial_t p_t + \nabla_x \cdot (v_t p_t) = 0, \tag{2.3}$$

in the weak sense, where

$$v_t(x) = \mathbb{E}[\dot\alpha(t)X_0 + \dot\sigma(t)X_1 \mid X_t = x]. \tag{2.4}$$

By using the stochastic reverse-time sampler we can traverse the marginals $p_t(x)$ from $t = 1$ to $t = 0$ and thereby generate new samples from the data distribution $p_0(x)$.

Proposition 2.3 (Reverse-time sampler, Albergo et al., 2025)**.** *Assume p_t is positive and sufficiently smooth, and that the SDE below is well posed and its Fokker–Planck equation has a unique probability-density solution. Define $s_t(x) = \nabla_x \log p_t(x)$ and let $w\colon [0,1] \to \mathbb{R}_{\geq 0}$ be continuous. With W_ℓ a standard n-dimensional Brownian motion, set*

$$\mathrm{d}X^{\mathrm{R}}_\ell = \big[-v_{1-\ell}(X^{\mathrm{R}}_\ell) + w(\ell)s_{1-\ell}(X^{\mathrm{R}}_\ell)\big]\mathrm{d}\ell + \sqrt{2w(\ell)}\,\mathrm{d}W_\ell, \quad X^{\mathrm{R}}_0 \sim \mathcal{N}(0,\Sigma). \tag{2.5}$$

Then X^{R}_ℓ has density $p_{1-\ell}$.

We consider the special case of $w = 0$, where (2.5) simplifies to the ODE,

$$\dot X_t = v_t(X_t), \qquad X_1 \sim \mathcal{N}(0,\Sigma), \tag{2.6}$$

integrated from $t = 1$ toward $t = 0$. For $m_t(x) := \mathbb{E}[X_0 \mid X_t = x]$ and $e_t(x) := \mathbb{E}[X_1 \mid X_t = x]$, the discretized interpolant update from t to $s < t$ is precisely noiseless DDIM (Song et al., 2020a):

$$\begin{aligned} x_s &= \alpha(s)m_t(x_t) + \sigma(s)e_t(x_t) \\ &= x_t + [\alpha(s) - \alpha(t)]m_t(x_t) + [\sigma(s) - \sigma(t)]e_t(x_t). \end{aligned} \tag{2.7}$$

Conditional sampling and classifier-free guidance. For a conditioning variable C, assume the noise X_1 is independent of (X_0, C). The conditional velocity is

$$v_t(x \mid c) = \mathbb{E}[\dot\alpha(t)X_0 + \dot\sigma(t)X_1 \mid X_t = x, C = c]. \tag{2.8}$$

The exact conditional flow $v_t(x \mid c)$ transports the conditional data distribution. However, in practice, classifier-free guidance is used to trade sample diversity for fidelity (Ho & Salimans, 2022), with guidance strength γ.

Definition 2.4 (Classifier-Free Guidance)**.** Given conditional flow $v_t(\cdot|c)$ and unconditional flow $v_t(\cdot)$, define, for guidance strength $\gamma \in \mathbb{R}$ the guided flow,

$$v_{\gamma,t}(x|c) = \gamma v_t(x|c) + (1-\gamma)v_t(x). \tag{2.9}$$

The choice $\gamma = 1$ recovers the original conditional flow, while $\gamma > 1$ extrapolates beyond it.

Prior work on guidance in toy low-dimensional settings from which practitioners have drawn inspiration (see e.g. Figure 3.1 of Zheng et al. (2023), reproduced here as the top row in Figure 1) assume that the unconditional distribution consists of the full data mixture while conditioning isolates a subset of features, i.e. the unconditional distribution $p(x)$ is given by $\sum_{k=1}^K \pi_k \mathcal{N}(\mu_k, \Sigma_k)$ for some parameters π_k, μ_k, Σ_k while the conditional distribution $p(x|c)$ is given by $\mathcal{N}(\mu_{j_c}, \Sigma_{j_c})$ for some j_c.

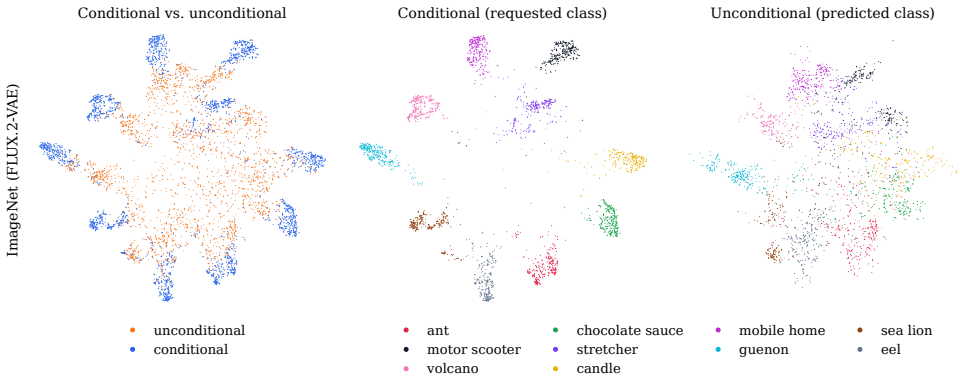

Figure 2: Empirical evidence suggests that the conditional ($\gamma = 1$) generation exhibits more distinctly class-specific samples of higher complexity relative to unconditional generation ($\gamma = 0$). We visualize t-SNE embeddings of DINO-V2 features of ImageNet samples using conditional vs. unconditional diffusion models for 10 different classes. Conditional samples are generated using class conditioning whereas unconditional samples are generated using rejection sampling with a ResNet-based classifier. See Figure 8 for more comparisons with different classes.

Motivation for our analytical model. Assuming the unconditional distribution is a more complex mixture would mean that the unconditional model is more expressive than the conditional model. As previously discussed, we believe this is not reflective of classifier-free guidance in practice, as (1) unconditional models exhibit far worse fidelity than their conditional counterparts (Ho & Salimans, 2022) and (2) our model captures techniques such as self-guidance (Li et al., 2025) and autoguidance (Karras et al., 2024), which use less-expressive conditional models in place of the "unconditional" flow. We follow Jin et al. (2026) in assuming that the unconditional flow is a "coarse" approximation and can be represented by a single Gaussian, while using a mixture to represent the conditional distribution. Figure 2 demonstrates precisely this structure via t-SNE embeddings of DINO-V2 features of generated ImageNet samples while Figure 1 contrasts our setup with that of prior work and highlights the magnification and mode-collapse effects that our exact form predicts.

3 Analytical Solutions to Guidance with Mixture-Based Conditional Distributions

We now present our main theoretical model and results, with proofs for this section deferred to Appendix C. Concretely, we model the unconditional distribution $p(x)$ by $\mathcal{N}(x; \bar{\mu}, \Sigma)$ and the conditional distribution $p(x|c)$ by a mixture of Gaussians which lie on a common-radius ellipse around $\bar{\mu}$. Throughout our analysis, we consider a fixed conditioning value c and write $p(x|c)$ to distinguish the conditional $\gamma = 1$ distribution from $p(x)$, the unconditional $\gamma = 0$ distribution.

Definition 3.1 (Common-radius Gaussian Mixture)**.** A common-radius Gaussian mixture has density,

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma), \qquad \pi_k > 0, \qquad \sum_{k=1}^{K} \pi_k = 1, \tag{3.1}$$

where $\Sigma \succ 0$ and there exists some center $\bar{\mu} \in \mathbb{R}^n$ and $R \geq 0$ such that $\|\mu_k - \bar{\mu}\|_{\Sigma^{-1}} = R$ for every k. We say that a conditional distribution $p(x|c)$ satisfies (3.1) at c if $p(\cdot|c)$ matches the above density.

Definition 3.2 (Common-radius Conditioning Property)**.** We say that for a fixed $C = c$, the densities $p(x|c)$ and $p(x)$ in X satisfy a common-radius conditioning property if, as a function of x, $p(x|c)$ has the form of Equation (3.1) for some $\mu_1, \ldots, \mu_k, \bar{\mu}$, and Σ and the unconditional density satisfies $p(x) = \mathcal{N}(x; \bar{\mu}, \Sigma)$.

Under Definition 3.2, the flow $v_t(x)$ for $p(x) = \mathcal{N}(x; \bar{\mu}, \Sigma)$ is simply $v_t(x) = \dot{\alpha}(t)\bar{\mu}$ (see Proposition C.1), while the conditional vector field $v_t(x|c)$ has the following special form:

Proposition 3.3 (Mixture velocity)**.** *Suppose that $X_0 \,|\, C = c$ is a common-radius mixture as in Definition 3.1 with center $\bar{\mu}$, means $\mu_1, \ldots, \mu_K$, and covariance Σ. Let $X_1 \sim \mathcal{N}(0, \Sigma)$ be*

independent of $X_0 \mid C = c$. *Then the velocity* $v_t(x|c)$ *defined in Definition 2.1 satisfies,*

$$v_t(x|c) = \dot{\alpha}(t) \sum_k w_k(t,x)\mu_k,$$
$$w_k(t,x) = \frac{\pi_k e^{\alpha(t)(x-\alpha(t)\bar{\mu})^\top \Sigma^{-1}(\mu_k - \bar{\mu})}}{\sum_j \pi_j e^{\alpha(t)(x-\alpha(t)\bar{\mu})^\top \Sigma^{-1}(\mu_j - \bar{\mu})}}. \tag{3.2}$$

Approach Sketch: The key property of Proposition 3.3 which enables our analysis is that the family of flows arising from Definition 3.1 is *closed* under scaling when $\bar{\mu} = 0$. Multiplying such a $v_t(x|c)$ by γ with $\bar{\mu} = 0$ results in correspondingly scaled means $\gamma\mu_k$ and covariances $\gamma\Sigma$. This can be directly extended to the setting where $\bar{\mu} \neq 0$ by considering the $\mu_k - \bar{\mu}$ offset instead. Given $\mu_1, \ldots, \mu_k$, $\bar{\mu}$, and Σ, we define the *auxiliary density* $\hat{p}_{\gamma,t}$,

$$\hat{p}_{\gamma,t}(x|c) := \sum_{k=1}^{K} \pi_k \mathcal{N}\big(x; \alpha(t)[\bar{\mu} + \gamma(\mu_k - \bar{\mu})], \gamma\Sigma\big). \tag{3.3}$$

The above $\hat{p}_{\gamma,t}$ satisfies the continuity equation (2.3) for $v_{\gamma,t}$ (i.e. $\partial_t \hat{p}_{\gamma,t} + \nabla_x \cdot (v_{\gamma,t}\hat{p}_{\gamma,t}) = 0$), but with a mismatched boundary condition. At $t = 1$, $\hat{p}_{\gamma,1}(x)$ is the γ-dependent $\mathcal{N}(0, \gamma\Sigma)$, instead of the desired $\mathcal{N}(0, \Sigma)$. Sampling using the standard reverse process (2.6) using the desired $p_1(x) = \mathcal{N}(x; 0, \Sigma)$ therefore reweights $\hat{p}_{t,\gamma}$ in a manner that Theorem 3.4 makes explicit:

Theorem 3.4 (Guided Density)**.** *Fix some* $C = c$ *for which the conditional density* $p_0(x|c)$ *satisfies Definition 3.1 for* $\mu_1, \ldots, \mu_K$, $\bar{\mu}$, *and* Σ *and let* $p_0(x) = \mathcal{N}(\bar{\mu}, \Sigma)$. *For any* $\gamma > 0$ *the guided velocity conditioned at* c *is given by,*

$$v_{\gamma,t}(x|c) = \dot{\alpha}(t) \left[\bar{\mu} + \gamma \sum_k w_k(t,x)(\mu_k - \bar{\mu}) \right]. \tag{3.4}$$

Define $y_{\gamma,t}$ *as a function which transports samples from* $\hat{p}_{\gamma,1}$ *to* $\hat{p}_{\gamma,t}$ *through* $v_{\gamma,t}$ *with corresponding inverse map* $y_{\gamma,t}^{-1}$, *i.e.*

$$y_{\gamma,t}(x) = z_t, \qquad \text{where} \qquad \dot{z}_s = v_{\gamma,s}(z_s|c), \qquad z_1 = x. \tag{3.5}$$

Then the unique family of probability densities $p_{\gamma,t}$ *satisfying both the transport equation for the guided velocity* $v_{\gamma,t}$ *in (3.4) and the boundary condition* $p_{\gamma,1} = \mathcal{N}(0, \Sigma)$ *is,*

$$p_{\gamma,t}(x|c) = \underbrace{\gamma^{n/2} \exp\left(-\frac{\gamma - 1}{2\gamma} \|y_{\gamma,t}^{-1}(x)\|^2_{\Sigma^{-1}} \right)}_{W_{\gamma,t}(x)} \hat{p}_{\gamma,t}(x|c), \tag{3.6}$$

where $\hat{p}_{\gamma,t}$ *is as defined in Equation (3.3). For* $\gamma > 1$, *since* $y_{\gamma,t}$ *is a diffeomorphism, the reweighting function* $W_{\gamma,t}$ *satisfies*

$$W_{\gamma,t}(y_{\gamma,t}(0)) = \gamma^{n/2} = \max_x W_{\gamma,t}(x), \qquad \lim_{\|x\|_{\Sigma^{-1}} \to \infty} W_{\gamma,t}(x) = 0. \tag{3.7}$$

Note that both the reweighting functions $W_{\gamma,t}(x)$ and $y_{\gamma,t}(x)$ are dependent on $v_{\gamma,t}(x|c)$ and therefore also implicitly depend on the fixed choice of c, although we omit this for notational convenience.

Interpretation: Equation (3.6) highlights two complementary effects of guidance for $\gamma > 1$:

(1) *Component-Specific Magnification*: The covariance Σ and component-specific features $\mu_k - \bar{\mu}$ in $\hat{p}_{\gamma,t}$ are scaled by γ.
(2) *Mode Collapse*: The resulting scaled distribution $\hat{p}_{t,\gamma}$ is then focused via $W_{t,\gamma}$ around $y_{\gamma,t}(0)$, the x_t associated with $x_1 = 0$ under the guided flow. As a result of this effect, $p_{\gamma,t}$ is skewed towards $y_{\gamma,t}(0)$.

These effects explain both (1) the exaggeration of class-specific features and (2) the loss of sample diversity under high guidance as $y_{\gamma,t}(0)$ is over-represented. Note that while (1) increases the data covariance, this effect is itself counteracted by the concentration (2), which depends on the specific placement of $\mu_1, \ldots, \mu_K$ and $\bar{\mu}$. In Section 4 we will show that both of these effects can be visualized in image diffusion models by considering projections along the predicted directions.

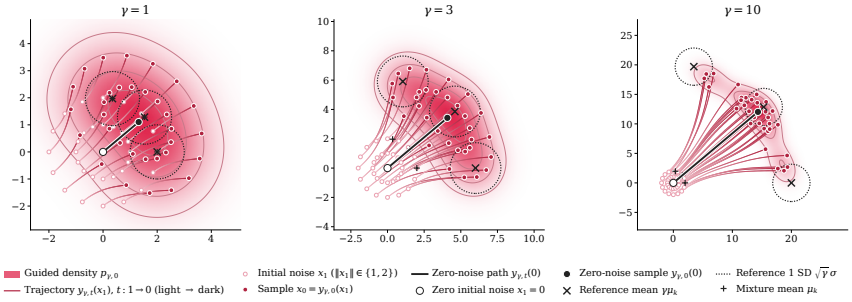

Figure 3: Guided flow map trajectories for a group of initial points while varying classifier-free guidance scales, derived from our analytical model in Section 3. Increasing guidance scale concentrates the guided distribution (shown as shaded density) towards the zero-noise sample $y_{\gamma,0}(0)$.

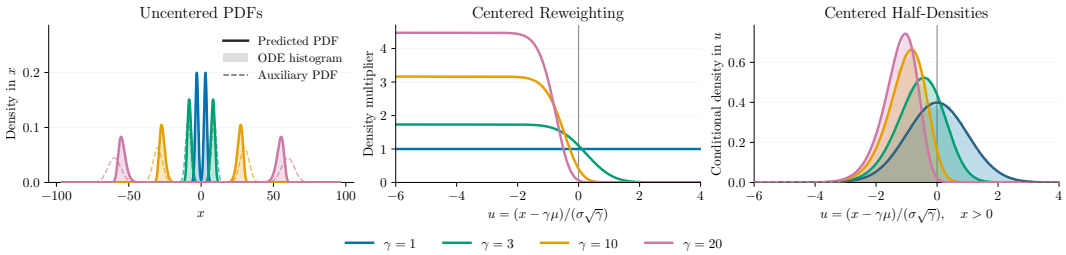

Figure 4: One-dimensional closed-forms for the $\gamma = 1$ mixture $\frac{1}{2}\mathcal{N}(-3, 1) + \frac{1}{2}\mathcal{N}(3, 1)$ and $\gamma = 0$ unconditional $\mathcal{N}(0, 1)$, for $\gamma \in \{1, 3, 10, 20\}$. From left to right: **(1)** the densities $\hat{p}_{\gamma,t}$ (dashed) and $p_{\gamma,t}$ (solid), **(2)** the density multiplier $W_{\gamma,0}$, in $(x - \gamma\mu)/(\sigma\sqrt{\gamma})$ coordinates, and **(3)** the positive-half densities of $p_{\gamma,0}$, also in rescaled coordinates. Lines use the exact forms in Corollary 3.5 while shaded areas use a histogram with 100,000 integrated samples.

Although Equation (3.6) offers insights into the structure of guided distributions, computing $y_{\gamma,t}$ still requires solving an ODE. In the following section we show that for particular distributions the dependency on $y_{\gamma,t}$ may be removed either in part or in its entirety, yielding the first exact closed-form solution for a guided density with a nonlinear flow.

3.1 Explicit Closed-Forms for the Guided Distribution

In one dimension the CDF of the auxiliary distribution $\hat{p}_{t,\gamma}$ is constant on $y_{\gamma,t}(x)$ for all t, which allows us to eliminate $y^{-1}_{\gamma,t}(x)$ from Equation (3.6) and write $p_{\gamma,t}(x|c)$ with $W_{\gamma,t}$ expressed in terms of the standard normal CDF Φ and its inverse Φ^{-1}.

Corollary 3.5 (Two-component Guided Density)**.** *Choose $\mu_1, \mu_2 \in \mathbb{R}$ satisfying $|\mu_1| = |\mu_2|$, and $\tau, \gamma > 0$. Let Φ denote the standard normal CDF. For $p(x|c) = \frac{1}{2}\mathcal{N}(x; \mu_1, \tau^2) + \frac{1}{2}\mathcal{N}(x; \mu_2, \tau^2)$ and $p(x) = \mathcal{N}(x; 0, \tau^2)$, define,*

$$F_{\gamma,t}(x) := \frac{1}{2}\sum_{k=1}^{2} \Phi\left(\frac{x - \gamma\alpha(t)\mu_k}{\tau\sqrt{\gamma}}\right), \quad W_{\gamma,t}(x) = \sqrt{\gamma}\exp\left(-\frac{\gamma - 1}{2}[\Phi^{-1}(F_{\gamma,t}(x))]^2\right). \tag{3.8}$$

Then the density $p_{\gamma,t}$ for $v_{\gamma,t}$ as defined in Equation (3.6) is given by,

$$p_{\gamma,t}(x|c) = \frac{W_{\gamma,t}(x)}{2}\sum_{k=1}^{2}\mathcal{N}(x; \gamma\alpha(t)\mu_k, \gamma\tau^2). \tag{3.9}$$

Interpretation: Corollary 3.5 demonstrates that for $\gamma > 1$ the density $p_{\gamma,t}$ is no longer a mixture of Gaussians, but rather a mixture of *skewed Gaussians* whose components have "centered" (using $u = x - \gamma\alpha(t)\mu_k$) densities of the form $g(u)\mathcal{N}(u; 0, \gamma\tau^2)$ and where the skewing function $g(u) = W_{\gamma,t}(u + \gamma\alpha(t)\mu_k)$ is a sigmoid-like function that tilts the components towards $y_{\gamma,t}(0)$. This suggests

that while guidance causes magnification of the $\mu_k - \bar{\mu}$ component, it also simultaneously suppresses the outer tails of the distribution. We visualize $\hat{p}, p$, and $W_{\gamma,t}$ for Corollary 3.5 in Figure 4.

Extension to Higher Dimensions: We can generalize the above result in a straightforward manner to higher dimensions, provided that the mixture components additionally have symmetries around the direction of interest. We state this result informally, to give intuition about the behavior of $W_{\gamma,t}$.

Theorem 3.6 (Guidance Under Symmetries, Informal, see Theorem C.3)**.** *For simplicity, assume $\bar{\mu} = 0$ and $\sum_k \pi_k \mu_k = 0$. For directions e whose ray is fixed by symmetries of the weighted component means (which we formalize in Definition C.2), the flow $v_{\gamma,t}(re)$ lies along e for all r and $W_{\gamma,t}$ satisfies,*

$$W_{\gamma,t}(re) = \gamma^{n/2} \exp\left(-\frac{\gamma-1}{2}[\Phi^{-1}(F_{\gamma,t}(r))]^2\right).$$

for the CDF $F_{\gamma,t}$ of a continuous scalar random variable analogous to $\hat{p}_{\gamma,t}$ in Corollary 3.5 arising from the projected flow $\langle v_{\gamma,t}(re), e\rangle$.

Interpretation: As $\gamma \to \infty$, the magnification effect separates the components of $\hat{p}_{\gamma,t}$ and analogously flattens the 1-dimensional CDF $F_{\gamma,t}$. Therefore along these symmetric directions we can interpret $W_{\gamma,t}$ as behaving analogously to a step function in the large-γ limit, with μ_k- and Σ-dependent transition boundaries.

3.2 Lower and upper bounds on the reweighting function in higher dimensions

The above results suggest that although $W_{\gamma,t}(x)$ skews components towards $y_{\gamma,t}(0)$, this effect may not cause unbounded collapse to the "center" point, $y_{\gamma,t}(0)$. This intuition is also reflected in the original definition of $W_{\gamma,t}(x)$ in Equation (3.6), as $W_{\gamma,t}(x) \propto \exp(-\frac{\gamma-1}{2\gamma}\|y^{-1}_{\gamma,t}(x)\|^2_{\Sigma^{-1}})$, for which the coefficient in the exponent has the limit $(\gamma-1)/(2\gamma) \to 1/2$ as $\gamma \to \infty$.

However, this naive analysis does not account for potentially γ-dependent behavior in $y^{-1}_{\gamma,t}$ itself. In this section we derive bounds for $\|y^{-1}(x)\|_{\Sigma^{-1}}$ that do not require solving for y^{-1} globally and do not impose any additional restrictions on the problem setting. In particular, for the following theorem, we use local sensitivity analysis of y^{-1} to both upper and lower bound $W_{\gamma,t}(x)$ with respect to $\|x - y_{\gamma,t}(0)\|_{\Sigma^{-1}}$, instead of $\|y^{-1}_{\gamma,t}(x)\|_{\Sigma^{-1}}$.

Theorem 3.7 (Central Enhancement and Tail Suppression)**.** *Assume the setting of Theorem 3.4 with $n \geq 1$ and $\gamma > 1$. Then, for every $t \in [0,1]$ and any $x \in \mathbb{R}^n$, the density factor $W_{\gamma,t}(x)$ satisfies,*

$$\exp\left(-\frac{\gamma-1}{2\gamma}\|x - y_{\gamma,t}(0)\|^2_{\Sigma^{-1}}\right) \leq \gamma^{-n/2} W_{\gamma,t}(x) \leq \exp\left(-\frac{\gamma-1}{2\gamma} e^{-\gamma R^2 \alpha(t)^2} \|x - y_{\gamma,t}(0)\|^2_{\Sigma^{-1}}\right).$$

Note that at $x = y_{\gamma,t}(0)$ these bounds match and recover that $W_{\gamma,t}(y_{\gamma,t}(0)) = \gamma^{n/2}$.

Interpretation: Theorem 3.7 suggests that $p_{\gamma,t}$ behaves like the product of the scaled mixture density $\hat{p}_{\gamma,t}$ with a density $q(x) \propto W_{\gamma,t}(x)$ with Gaussian density bounds with respect to $\|\cdot\|_{\Sigma^{-1}}$ for which the "variance" $\hat{\sigma}^2$ is in the range,

$$1 \leq \frac{\gamma}{\gamma-1} \leq \hat{\sigma}^2 \leq \frac{\gamma}{\gamma-1} e^{\gamma R^2},$$

for $\gamma > 1$ (note that $q(x)$ is not necessarily zero-mean and so is not sub-Gaussian in the technical sense). The lower bound $\hat{\sigma} \geq 1$ demonstrates that while $W_{\gamma,t}$ concentrates $p_{\gamma,t}$ around $y_{\gamma,t}(0)$, **the degree of distribution collapse has a γ-independent lower bound**. This intuition also matches our exact results in Corollary 3.5 and Theorem 3.6, as the derived $W_{\gamma,t}$ do not collapse as $\gamma \to \infty$.

Note that the bound $\hat{\sigma} \geq 1$ is expressed in terms of the shared $\|\cdot\|_{\Sigma^{-1}}$ metric. For the standard Euclidean metric $\|\cdot\|$ any equivalent bound would be data-dependent and significantly more complex when the covariance Σ is not identical for all mixture components.

4 Experiments

The existence of *feature magnification* and *mode-collapse* effects due to guidance have already been widely documented in prior literature (Jin et al., 2026; Li et al., 2026). In these experiments, we

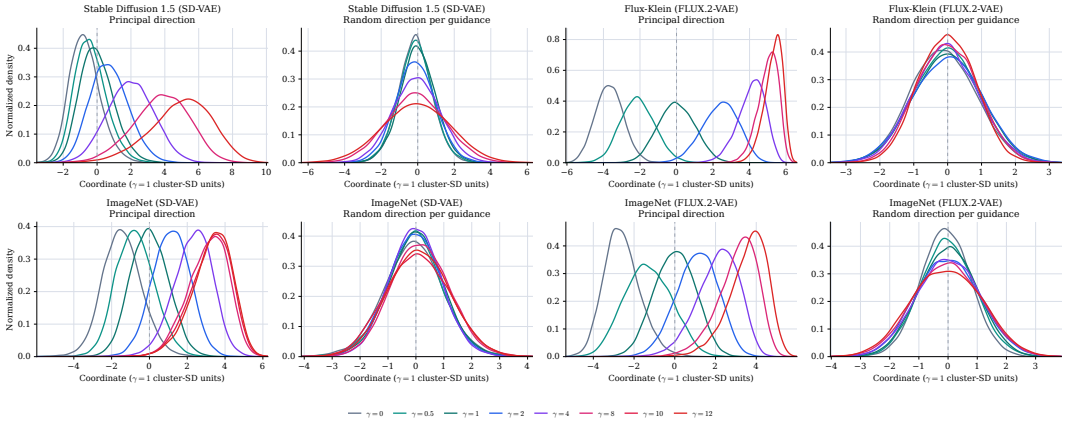

Figure 5: Principal direction and random-direction control densities across guidance scales. Each panel pairs a projection along the direction from the unconditional mean at $\gamma = 0$ to the cluster mean at $\gamma = 12$ with projection onto a random orthogonal direction (sampled independently per γ value). At each guidance value, within-cluster standardized shapes are mapped to a shared mean and variance and aggregated across both conditioning values and clusters. Coordinates are expressed in $\gamma = 1$ cluster standard-deviation units.

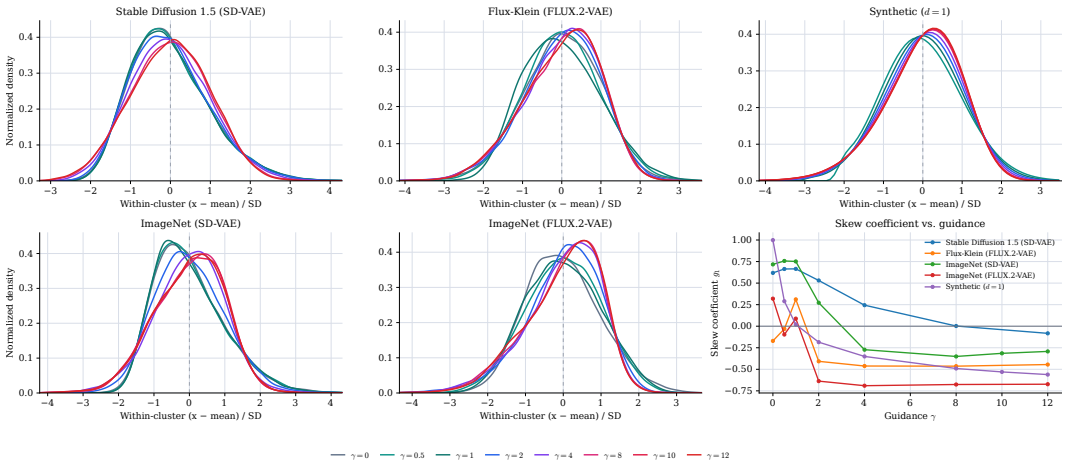

Figure 6: Within-cluster standardized densities across guidance, using a per-γ radial direction from each conditioning value's mean to each frozen cluster's mean. At each guidance value, each cluster's projections are normalized to zero mean and unit standard deviation before aggregation. The synthetic panel visualizes standardized distributions from the same setup as Figure 4.

demonstrate that their effects are visible precisely along our predicted *component-specific* axes and confirm that the degree of mode collapse is bounded, as suggested by our theory.

We conduct an empirical investigation into the behavior of latent image diffusion models in both text- and class-conditioned as well as different latent spaces. We use Stable Diffusion 1.5 and FLUX-Klein-9B for text-conditioned experiments and pretrained DiT models for class-conditional experiments on ImageNet. We use either noiseless DDIM or Euler-integration configurations to match the flow-based sampling ODE we analyze. See Appendix D for detailed experiment setup information.

For our experiments, we approximate $\bar{\mu}$ by the average over $\gamma = 0$ samples (i.e. $\bar{\mu} \approx \mathbb{E}_{p_{\gamma=0}(x)}[x]$). For $y_{\gamma,0}(0)$, we find that, with certain prompts and models, evaluating $y_{\gamma,0}(0)$ produces highly out-of-distribution images with significant artifacts and so use the average of $x \sim p_{\gamma,0}(x|c)$ as a proxy (see Figure 9).



Figure 7: Outputs across guidance scales $\gamma = 1, 4, 12$ for a selected ImageNet text prompt using FLUX latents (top) and FLUX-Klein-9B (bottom). The bottom row shows $y_{\gamma,0}(0)$ while the top row shows samples matched across guidance scales by the initial value for x_1, demonstrating that increasing guidance concentrates samples around images with features similar to $y_{\gamma,0}(0)$, while some variation persists.

Note that for our analysis to hold we do *not* require that $v_{\gamma,t}(y_{\gamma,t}(0))$ be well approximated by the learned model. The guided sampling procedure uses $x_1 \sim \mathcal{N}(0, I)$, which are in-distribution for the learned model, unlike $x_1 = 0$, for which one cannot necessarily expect a well-formed sample in any case. Our model predicts collapse to the true $y_{\gamma,t}(0)$ given the $p(x)$ and $p(x|c)$ distributions, *even if* the learned $v_{\gamma,t}$ poorly captures that specific flow.

We compute these metrics in the latent space and use K-means clustering, with $K \in \{1, \ldots, 16\}$ chosen via Bayesian Information Criterion, to compute $\{\mu_k\}_{k=1}^{K}$ for $p(x|c)$. This allows us to empirically investigate our hypothesized guidance effects:

1. In Figure 5 we project each cluster along the direction from the estimated unconditional mean at $\gamma = 0$ to the cluster mean at $\gamma = 12$. We show that higher guidance values push samples along this axis, while randomly chosen axes see no net effect, precisely what our model predicts.
2. In Figure 6 we plot the same clusters, projected along the $\mu_k - \mathbb{E}_{p_{\gamma,0}(x|c)}[x]$ direction, with clusters normalized to mean zero and unit standard deviation. We can observe that along this axis large γ induces left-skewed distributions, matching the predictions for the behavior of the reweighting function $W_{\gamma,t}(x)$.
3. In Figure 7 we visualize distributional collapse to a "prototypical" $y_{\gamma,0}(0)$ sample. We can clearly observe both concentration around $y_{\gamma,0}(0)$ in addition to persistence of other modes. For models using SD-VAE latents, $y_{\gamma,0}(0)$ produced highly artifacted samples (Figure 9), which we attribute to the interaction of ϵ-parameterization with the out-of-distribution $x_1 = 0$. As such, we only show FLUX-based models, which use velocity-prediction. For additional classes and samples, see Figure 11 and Figure 12.

Together, these results support our theoretical account of guidance as combining directional magnification with inward reweighting, producing concentration without eliminating all distinct modes.

5 Discussion and Conclusion

Our results suggest that guidance can be understood as amplifying cluster-specific features and amplifying a specific "central" point. As our results are predicated on the assumption that the unconditional flow is a poor representation of the true unconditional distribution, this opens the possibility that advances in unconditional generation may conversely reduce the benefits of classifier-free guidance. For limitations of our model and extended discussion, see Appendix B.

We believe our results should be viewed as complementary to existing results for linear flows, where classifier-free guidance yields a Gaussian with an altered mean and covariance. More broadly, an understanding of flow composition beyond the standard guidance scheme may allow the development of sampling methods with more fine-grained control of the sampling distribution.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

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.

Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

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

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Huayu Chen, Kai Jiang, Kaiwen Zheng, Jianfei Chen, Hang Su, and Jun Zhu. Visual generation without guidance. *arXiv preprint arXiv:2501.15420*, 2025a.

Huayu Chen, Hang Su, Peize Sun, and Jun Zhu. Toward guidance-free ar visual generation via condition contrastive alignment. In *International Conference on Learning Representations*, volume 2025, pp. 65862–65881, 2025b.

Muthu Chidambaram, Khashayar Gatmiry, Sitan Chen, Holden Lee, and Jianfeng Lu. What does guidance do? a fine-grained analysis in a simple setting. *Advances in Neural Information Processing Systems*, 37:84968–85005, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Hang Ding, Xue Wang, Tian Zhou, and Tao Yao. Simdiff: Simpler yet better diffusion model for time series point forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20781–20789, 2026.

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.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *Advances in Neural Information Processing Systems*, 38:75460–75482, 2026.

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. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Enze Jiang and Zheng Ma. Analytic distribution of classifier-free guidance for schedule design. *arXiv preprint arXiv:2607.19725*, 2026.

Cheng Jin, Qitan Shi, and Yuantao Gu. Stage-wise dynamics of classifier-free guidance in diffusion models. In *International Conference on Learning Representations*, volume 2026, pp. 154028–154053, 2026.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

Tiancheng Li, Weijian Luo, Zhiyang Chen, Liyuan Ma, and Guo-Jun Qi. Self-guidance: Boosting flow and diffusion generation on their own. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Xiang Li, Rongrong Wang, and Qing Qu. Towards understanding the mechanisms of classifier-free guidance. *Advances in Neural Information Processing Systems*, 38:62835–62889, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14297–14306. IEEE, 2023.

Hunter Nisonoff, Junhao Xiong, Stephan Allenspach, and Jennifer Listgarten. Unlocking guidance for discrete state-space diffusion and flow models. In *International Conference on Learning Representations*, volume 2025, pp. 36052–36106, 2025.

Krunoslav Lehman Pavasovic, Jakob Verbeek, Giulio Biroli, and Marc Mezard. Classifier-free guidance: From high-dimensional analysis to generalized guidance forms. *arXiv preprint arXiv:2502.07849*, 2025.

Daniel Pfrommer, Zehao Dou, Christopher Scarvelis, Max Simchowitz, and Ali Jadbabaie. Is your diffusion model actually denoising? *Advances in Neural Information Processing Systems*, 38: 94380–94420, 2026.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. Stay on topic with classifier-free guidance. *arXiv preprint arXiv:2306.17806*, 2023.

Christopher Scarvelis, Haitz Sáez de Ocáriz Borde, and Justin Solomon. Closed-form diffusion models. *arXiv preprint arXiv:2310.12395*, 2023.

Marta Skreta, Tara Akhound-Sadegh, Viktor Ohanesian, Roberto Bondesan, Alán Aspuru-Guzik, Arnaud Doucet, Rob Brekelmans, Alexander Tong, and Kirill Neklyudov. Feynman-kac correctors in diffusion: Annealing, guidance, and product of experts. *arXiv preprint arXiv:2503.02819*, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. *arXiv preprint arXiv:2208.06193*, 2022.

Yuchen Wu, Minshuo Chen, Zihao Li, Mengdi Wang, and Yuting Wei. Theoretical insights for diffusion guidance: A case study for gaussian mixture models. *arXiv preprint arXiv:2403.01639*, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Yanchun Yu, Weibin Zhang, and Yun Deng. Frechet inception distance (fid) for evaluating gans. *China University of Mining Technology Beijing Graduate School*, 3(11), 2021.

Kaiyu Yue, Menglin Jia, Ji Hou, and Tom Goldstein. Image generation with a sphere encoder. *arXiv preprint arXiv:2602.15030*, 2026.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 35791–35820, 2026.

Candi Zheng and Yuan Lan. Characteristic guidance: Non-linear correction for diffusion model at large guidance scale. *arXiv preprint arXiv:2312.07586*, 2023.

Qinqing Zheng, Matt Le, Neta Shaul, Yaron Lipman, Aditya Grover, and Ricky TQ Chen. Guided flows for generative modeling and decision making. *arXiv preprint arXiv:2311.13443*, 2023.

A Extended Prior Work

Guidance and the tilted-distribution interpretation. In its original formulation, diffusion (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020b) was proposed as a mechanism for generating unconditional samples by reversing a stochastic random walk to iteratively "denoise" random data. The development of deterministic samplers such as DDIM (Song et al., 2020a) and more general theoretical frameworks such as flow-map matching and stochastic interpolants (Albergo & Vanden-Eijnden, 2022; Albergo et al., 2025) has given diffusion a solid theoretical and empirical underpinning in this setting. However, prior to the introduction of *classifier guidance* by Dhariwal & Nichol (2021), naive extension of this framework to the conditional setting underperformed contemporaneous GAN-based alternatives such as Brock et al. (2018). Classifier guidance combines the scaled gradient of a noise-aware classifier with the output of a diffusion model with the aim of sampling from a density proportional to $p(x)\,p(c \mid x)^{\gamma}$. The subsequent development of *classifier-free guidance* (Ho & Salimans, 2022) removed the dependency on an external classifier by extrapolating between a conditional and an unconditional diffusion model instead, implicitly targeting the *tilted density* $p(x|c)^{\gamma}p(x)^{1-\gamma}$. For simplicity, we use *guidance* to refer primarily to the classifier-free variant of guidance throughout the rest of this work.

The fidelity–diversity trade-off of guidance has been documented since inception: FID (Yu et al., 2021), which measures feature-distribution-level similarity, is optimal at small guidance weights while Inception Score (Salimans et al., 2016), which more indirectly captures individual sample quality, keeps rising (Dhariwal & Nichol, 2021; Ho & Salimans, 2022), and large guidance weights γ drive pixel values to saturation (Saharia et al., 2022). As a result of this phenomenon, there has been considerable interest in understanding the precise effect of guidance on the resulting sample distribution.

Guidance does not sample the tilted distribution. It was first observed by Du et al. (2023) that guidance does not faithfully sample from the tilted distribution. Although guidance targets the distribution $p(x|c)^{\gamma}p(x)^{1-\gamma}$ at each sampling step, the product of noised $p(x|c)$ and $p(x)$ distributions does not match the density of a noised $p(x|c)^{\gamma}p(x)^{1-\gamma}$ distribution, leading to intrinsic bias. Bradley & Nakkiran (2024) show that guided DDPM and guided DDIM produce different distributions, neither of which is the γ-tilted density, and that in the SDE limit guidance acts as a predictor for the conditional distribution alternating with a Langevin corrector for the tilted one. Chidambaram et al. (2024) prove that, for mixtures of compactly supported components, increasing the guidance weight concentrates mass on the boundary of the conditional support rather than on its modes. Recent analysis by Jiang & Ma (2026) write the distribution induced by the guided probability-flow ODE as a path-integral correction to the marginal. Each of these expresses the guided density through a correction accumulated along the sampling trajectory, rather than in closed form.

Gaussian and Gaussian-mixture analyses. The closest antecedents to this work analyze guidance with exact scores on Gaussian or Gaussian-mixture data. Wu et al. (2024) prove formally that guidance increases classification confidence and decreases the differential entropy of the output for Gaussian mixtures. Li et al. (2026) decompose guidance with linear flows arising from a Gaussian model into a mean shift towards the class-specific mean together with contrastive principal-component terms that amplify class-specific directions and suppress directions shared with the unconditional distribution while Jin et al. (2026) analyze guidance effects across different time regimes. Pavasovic et al. (2025) characterize the guided distribution for a two-cluster mixture in high dimension, finding an overshot mean and shrunken variance at finite dimension that vanish as $d \to \infty$.

These results provide monotonicity statements or stage-wise descriptions, but none gives an exact, finite-dimensional density for the nonlinear flows that arise from multimodal conditional distributions. Our restriction to mixtures with a common Mahalanobis radius is what makes such a formula directly available.

Guidance beyond continuous diffusion. The same guidance-style extrapolation is now used well outside image diffusion, which underlines the value of understanding its effect on the sampled distribution. It improves prompt adherence in autoregressive language models (Sanchez et al., 2023), transfers directly to flow matching (Zheng et al., 2023), and is standard in masked and autoregressive image generation (Chang et al., 2023), where it can also be absorbed into training (Chen et al.,

2025a;b). Novel one-step generative methods, such as MeanFlow Geng et al. (2026) and Sphere Encoders (Yue et al., 2026) have devised guidance-style schemes for boosting empirical performance. Finally, guided samplers are routinely distilled into single networks conditioned on the guidance weight (Meng et al., 2023), so that any systematic skew introduced by guidance is inherited by the distilled model.

B Limitations and Extended Discussion

Our analysis relies on a particular model for both the unconditional and conditional distributions. It additionally requires the unconditional and conditional components to share the same covariance matrix, which is an unrealistic assumption, as we would expect the unconditional distribution to exhibit significantly higher covariance. Thus, we view our model as orthogonal to the intuition provided by existing result involving linear flows (Li et al., 2026; Wu et al., 2024), which can handle mismatched covariances between the unconditional and conditional distributions. With more precise analysis under a more general class of distributions, it may be possible to remove some of these structural limitations for our work.

Our results also presume that the learned diffusion model exactly matches the ground-truth flow for the modeled distributions. It does not capture effects such as generalization of the model itself to unobserved conditioning values, which may introduce its own biases (Pfrommer et al., 2026; Scarvelis et al., 2023). Results in the generalization of diffusion models suggest that models may generalize to unseen conditioning values via a form of guidance analogous to classifier-free guidance, meaning that our results may provide insight into generalization behavior, although we treat such analysis as out-of-scope for this work.

Finally, it is unclear if our proposed model for guidance will apply given future advancements in the fidelity of samples generated by unconditional diffusion models or flows. It may be that for more expressive unconditional flows guidance does not confer fidelity benefits (as a more powerful unconditional model may invert the self-guidance effect observed in Li et al. (2025)). This may effectively serve as a litmus test as to whether our model captures the aspects of guidance that are relevant to the improved sample quality it currently produces.

C Deferred proofs

Proposition C.1 (Finite Gaussian mixture). *Fix $\mu_1, \dots, \mu_K \in \mathbb{R}^d$, covariances $\Sigma_1, \dots, \Sigma_K \succ 0$, and some schedule (α, σ). Let X_0 be distributed according to,*

$$X_0 \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k), \qquad \pi_k > 0, \qquad \sum_{k=1}^{K} \pi_k = 1, \tag{C.1}$$

and $X_1 \sim \mathcal{N}(0, \Sigma)$. Define $\Sigma_{k,t} := \alpha(t)^2 \Sigma_k + \sigma(t)^2 \Sigma$. Then the density $p_t(x)$ of the interpolant X_t and associated v_t are given by

$$p_t(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}\big(x; \alpha(t)\mu_k, \Sigma_{k,t}\big), \tag{C.2}$$

$$v_t(x) = \sum_{k} w_k(t,x) \left[\dot\alpha(t)\mu_k + (\alpha(t)\dot\alpha(t)\Sigma_k + \sigma(t)\dot\sigma(t)\Sigma)\Sigma_{k,t}^{-1}(x - \alpha(t)\mu_k)\right]. \tag{C.3}$$

where

$$w_k(t,x) := \frac{\frac{\pi_k}{\sqrt{\det(\Sigma_{k,t})}} e^{-\frac{1}{2}\|x-\alpha(t)\mu_k\|^2_{\Sigma_{k,t}^{-1}}}}{\sum_j \frac{\pi_j}{\sqrt{\det(\Sigma_{j,t})}} e^{-\frac{1}{2}\|x-\alpha(t)\mu_j\|^2_{\Sigma_{j,t}^{-1}}}}.$$

Proof. By Equation (2.4),

$$v_t(x) = \mathbb{E}[\dot\alpha(t)X_0 + \dot\sigma(t)X_1 | X_t = x]$$

$$= \mathbb{E}\left[\dot{\alpha}(t)X_0 + \frac{\dot{\sigma}(t)}{\sigma(t)}(X_t - \alpha(t)X_0)\middle| X_t = x\right]$$
$$= \frac{\dot{\sigma}(t)}{\sigma(t)}x + \left(\frac{\dot{\alpha}(t)\sigma(t) - \alpha(t)\dot{\sigma}(t)}{\sigma(t)}\right)\mathbb{E}[X_0|X_t = x]$$

Let C be drawn from the categorical distribution over labels and X_0 be distributed such that $X_0|C \overset{\text{dist}}{=} \mathcal{N}(\mu_k, \Sigma_k)$. We can then write,

$$\begin{aligned}
\mathbb{E}[X_0|X_t = x] &= \sum_{k=1}^{K} P(C = k|X_t = x)\mathbb{E}[X_0|X_t = x, C = k] \\
&= \sum_{k=1}^{K} w_k(t,x)\int y\left(\frac{p(X_t = x|X_0 = y, C = k)p(X_0 = y|C = k)}{p(X_t = x|C = k)}\right)\mathrm{d}y \\
&= \sum_{k=1}^{K} w_k(t,x)\int yC_k \exp\left(-\frac{1}{2\sigma(t)^2}\|x - \alpha(t)y\|^2_{\Sigma^{-1}} - \frac{1}{2}\|y - \mu_k\|^2_{\Sigma_k^{-1}} + \frac{1}{2}\|x - \alpha(t)\mu_k\|^2_{\Sigma_{k,t}^{-1}}\right)\mathrm{d}y \\
&= \sum_{k=1}^{K} w_k(t,x)\int yC'_k(x) \exp\left(-\frac{\alpha(t)^2}{2\sigma(t)^2}\left\|\frac{1}{\alpha(t)}x - y\right\|^2_{\Sigma^{-1}} - \frac{1}{2}\|y - \mu_k\|^2_{\Sigma_k^{-1}}\right)\mathrm{d}y \\
&= \sum_{k=1}^{K} w_k(t,x)\int yC'_k(x) \exp\Big(-\frac{1}{2}(\alpha(t)^{-1}x - y)^\top\left(\frac{\sigma(t)^2}{\alpha(t)^2}\Sigma\right)^{-1}(\alpha(t)^{-1}x - y) \\
&\qquad\qquad - \frac{1}{2}(y - \mu_k)^\top\Sigma_k^{-1}(y - \mu_k)\Big)\mathrm{d}y
\end{aligned}$$

The product of two Gaussian densities $\mathcal{N}(x;\mu_A,\Sigma_A)$ and $\mathcal{N}(x;\mu_B,\Sigma_B)$ is proportional to the density$\mathcal{N}(x;(\Sigma_A^{-1}+\Sigma_B^{-1})^{-1}(\Sigma_A^{-1}\mu_A+\Sigma_B^{-1}\mu_B),(\Sigma_A^{-1}+\Sigma_B^{-1})^{-1})$. In this case the normalization constant of the resulting density is precisely $C'_k(x)$, meaning we are integrating over a density with mean $(\sigma(t)^2\Sigma_k^{-1} + \alpha(t)^2\Sigma^{-1})^{-1}(\sigma(t)^2\Sigma_k^{-1}\mu_k + \alpha(t)\Sigma^{-1}x)$.

$$\begin{aligned}
&= \sum_{k=1}^{K} w_k(t,x)\left(\Sigma_k^{-1} + \frac{\alpha(t)^2}{\sigma(t)^2}\Sigma^{-1}\right)^{-1}(\Sigma_k^{-1}\mu_k + \frac{\alpha(t)}{\sigma(t)^2}\Sigma^{-1}x) \\
&= \sum_{k=1}^{K} w_k(t,x)\left(\sigma(t)^2\Sigma_k^{-1} + \alpha(t)^2\Sigma^{-1}\right)^{-1}(\sigma(t)^2\Sigma_k^{-1}\mu_k + \alpha(t)\Sigma^{-1}x)
\end{aligned}$$

Substituting into the original expression, we have

$$v_t(x) = \sum_k w_k(t,x)\left[\dot{\alpha}(t)\mu_k + (\alpha(t)\dot{\alpha}(t)\Sigma_k + \sigma(t)\dot{\sigma}(t)\Sigma)\Sigma_{k,t}^{-1}(x - \alpha(t)\mu_k)\right]. \tag{C.4}$$

□

Proposition 3.3 (Mixture velocity)**.** *Suppose that $X_0 \,|\, C = c$ is a common-radius mixture as in Definition 3.1 with center $\bar{\mu}$, means $\mu_1, \ldots, \mu_K$, and covariance Σ. Let $X_1 \sim \mathcal{N}(0, \Sigma)$ be independent of $X_0 \,|\, C = c$. Then the velocity $v_t(x|c)$ defined in Definition 2.1 satisfies,*

$$\begin{aligned}
v_t(x|c) &= \dot{\alpha}(t)\sum_k w_k(t,x)\mu_k, \\
w_k(t,x) &= \frac{\pi_k e^{\alpha(t)(x-\alpha(t)\bar{\mu})^\top\Sigma^{-1}(\mu_k-\bar{\mu})}}{\sum_j \pi_j e^{\alpha(t)(x-\alpha(t)\bar{\mu})^\top\Sigma^{-1}(\mu_j-\bar{\mu})}}.
\end{aligned} \tag{3.2}$$

Proof. This is just a direct specialization of Proposition C.1. Since $\Sigma_{k,t} = \Sigma$ and $\sigma^2(t) + \alpha^2(t) = 1$, $\sigma(t)\dot{\sigma}(t) = -\alpha(t)\dot{\alpha}(t)$, so the expression in (C.4) simplifies to

$$v_t(x) = \dot{\alpha}(t)\sum_k w_k(t,x)\mu_k,$$

and the expression for the weights $w_k(t,x)$ follows by specializing to the case where $(\mu_k - \bar{\mu})^\top\Sigma^{-1}(\mu_k - \bar{\mu}) = R^2$. □

Theorem 3.4 (Guided Density). *Fix some $C = c$ for which the conditional density $p_0(x|c)$ satisfies Definition 3.1 for $\mu_1, \dots, \mu_K$, $\bar{\mu}$, and Σ and let $p_0(x) = \mathcal{N}(\bar{\mu}, \Sigma)$. For any $\gamma > 0$ the guided velocity conditioned at c is given by,*

$$v_{\gamma,t}(x|c) = \dot{\alpha}(t)\left[\bar{\mu} + \gamma \sum_k w_k(t,x)(\mu_k - \bar{\mu})\right]. \tag{3.4}$$

Define $y_{\gamma,t}$ as a function which transports samples from $\hat{p}_{\gamma,1}$ to $\hat{p}_{\gamma,t}$ through $v_{\gamma,t}$ with corresponding inverse map $y_{\gamma,t}^{-1}$, i.e.

$$y_{\gamma,t}(x) = z_t, \qquad \text{where} \qquad \dot{z}_s = v_{\gamma,s}(z_s|c), \qquad z_1 = x. \tag{3.5}$$

Then the unique family of probability densities $p_{\gamma,t}$ satisfying both the transport equation for the guided velocity $v_{\gamma,t}$ in (3.4) and the boundary condition $p_{\gamma,1} = \mathcal{N}(0, \Sigma)$ is,

$$p_{\gamma,t}(x|c) = \underbrace{\gamma^{n/2} \exp\left(-\frac{\gamma - 1}{2\gamma} \|y_{\gamma,t}^{-1}(x)\|_{\Sigma^{-1}}^2\right)}_{W_{\gamma,t}(x)} \hat{p}_{\gamma,t}(x|c), \tag{3.6}$$

where $\hat{p}_{\gamma,t}$ is as defined in Equation (3.3). For $\gamma > 1$, since $y_{\gamma,t}$ is a diffeomorphism, the reweighting function $W_{\gamma,t}$ satisfies

$$W_{\gamma,t}(y_{\gamma,t}(0)) = \gamma^{n/2} = \max_x W_{\gamma,t}(x), \qquad \lim_{\|x\|_{\Sigma^{-1}} \to \infty} W_{\gamma,t}(x) = 0. \tag{3.7}$$

Proof. By Proposition 3.3, $v_{\gamma,t}(x)$ satisfies $\partial_t \hat{p}_{\gamma,t} + \nabla_x \cdot (v_{\gamma,t} \hat{p}_{\gamma,t}) = 0$ for

$$\hat{p}_{\gamma,t}(x) := \sum_{k=1}^{K} \pi_k \mathcal{N}\big(x; \alpha(t)[\bar{\mu} + \gamma(\mu_k - \bar{\mu})], \gamma\Sigma\big).$$

Indeed, applying that proposition with means $\bar{\mu} + \gamma(\mu_k - \bar{\mu})$, center $\bar{\mu}$, and covariance $\gamma\Sigma$ leaves the weights $w_k(t,x)$ unchanged and gives (3.4). The auxiliary density has the boundary condition $\hat{p}_{\gamma,1}(x) = \mathcal{N}(x; 0, \gamma\Sigma)$.

Define $W_{\gamma,t}(x) := \frac{p_{\gamma,1}(y_{\gamma,t}^{-1}(x))}{\hat{p}_{\gamma,1}(y_{\gamma,t}^{-1}(x))}$ and $p_{\gamma,t}(x) := W_{\gamma,t}(x)\hat{p}_{\gamma,t}(x)$, such that $p_{\gamma,1}(x) = \mathcal{N}(x; 0, \Sigma)$. Taking the ratio of the two Gaussian densities gives

$$W_{\gamma,t}(x) = \gamma^{n/2} \exp\left(-\frac{\gamma - 1}{2\gamma} \|y_{\gamma,t}^{-1}(x)\|_{\Sigma^{-1}}^2\right).$$

Since $y_{\gamma,1}$ is the identity, the reweighted $p_{\gamma,t}$ satisfies the required boundary condition. To show that it still satisfies the continuity equation, write

$$\partial_t p_{\gamma,t}(x) = \partial_t W_{\gamma,t}(x)\hat{p}_{\gamma,t}(x) + W_{\gamma,t}(x)\partial_t \hat{p}_{\gamma,t}(x) \tag{C.5}$$

$$= \partial_t W_{\gamma,t}(x)\hat{p}_{\gamma,t}(x) - W_{\gamma,t}(x)\nabla_x \cdot (\hat{p}_{\gamma,t} v_{\gamma,t}). \tag{C.6}$$

Note that by cancellation of $y_{\gamma,t}$ and $y_{\gamma,t}^{-1}$, $W_{\gamma,t}(y_{\gamma,t}(x))$ is constant in t, so

$$0 = \frac{\mathrm{d}}{\mathrm{d}t} W_{\gamma,t}(y_{\gamma,t}(x)) = \partial_t W_{\gamma,t}(y_{\gamma,t}(x)) + \nabla_x W_{\gamma,t}(y_{\gamma,t}(x)) \cdot v_{\gamma,t}(y_{\gamma,t}(x)).$$

Since $y_{\gamma,t}$ is onto, $\partial_t W_{\gamma,t}(x) = -\nabla_x W_{\gamma,t}(x) \cdot v_{\gamma,t}(x)$ holds for all $x \in \mathbb{R}^n, t \in [0,1]$. Substituting this into (C.6) and recognizing the product rule expression yields,

$$\partial_t p_{\gamma,t} = -\hat{p}_{\gamma,t}\, v_{\gamma,t} \cdot \nabla_x W_{\gamma,t} - W_{\gamma,t}\nabla_x \cdot (\hat{p}_{\gamma,t} v_{\gamma,t}) = -\nabla_x \cdot (p_{\gamma,t} v_{\gamma,t}).$$

Lastly, since normalization is preserved by the transport equation, since $p_{\gamma,1}(x) = \mathcal{N}(x; 0, \Sigma)$, $p_{\gamma,t}(x)$ is a normalized density for all $t \in [0,1]$. □

Corollary 3.5 (Two-component Guided Density). *Choose $\mu_1, \mu_2 \in \mathbb{R}$ satisfying $|\mu_1| = |\mu_2|$, and $\tau, \gamma > 0$. Let Φ denote the standard normal CDF. For $p(x|c) = \frac{1}{2}\mathcal{N}(x; \mu_1, \tau^2) + \frac{1}{2}\mathcal{N}(x; \mu_2, \tau^2)$ and $p(x) = \mathcal{N}(x; 0, \tau^2)$, define,*

$$F_{\gamma,t}(x) := \frac{1}{2}\sum_{k=1}^{2} \Phi\left(\frac{x - \gamma\alpha(t)\mu_k}{\tau\sqrt{\gamma}}\right), \quad W_{\gamma,t}(x) = \sqrt{\gamma} \exp\left(-\frac{\gamma - 1}{2}[\Phi^{-1}(F_{\gamma,t}(x))]^2\right). \tag{3.8}$$

Then the density $p_{\gamma,t}$ for $v_{\gamma,t}$ as defined in Equation (3.6) is given by,

$$p_{\gamma,t}(x|c) = \frac{W_{\gamma,t}(x)}{2} \sum_{k=1}^{2} \mathcal{N}(x; \gamma\alpha(t)\mu_k, \gamma\tau^2). \tag{3.9}$$

Proof. Note that $\hat{p}_{\gamma,t}(x) = \frac{1}{2}\mathcal{N}(x; \gamma\alpha(t)\mu_1, \gamma\tau^2) + \frac{1}{2}\mathcal{N}(x; \gamma\alpha(t)\mu_2, \gamma\tau^2)$ satisfies the continuity equation for $v_{\gamma,t}(x)$. Integrating the continuity equation yields

$$\begin{aligned} 0 &= \int_{-\infty}^{x} [\partial_t \hat{p}_{\gamma,t}(u) + \partial_u(\hat{p}_{\gamma,t}(u)v_{\gamma,t}(u))]\,\mathrm{d}u \\ &= \partial_t F_{\gamma,t}(x) + \hat{p}_{\gamma,t}(x)v_{\gamma,t}(x) \\ &= \partial_t F_{\gamma,t}(x) + [\partial_x F_{\gamma,t}(x)]v_{\gamma,t}(x). \end{aligned}$$

Thus for any given x, $\frac{\mathrm{d}}{\mathrm{d}t}F_{\gamma,t}(y_{\gamma,t}(x)) = \partial_t F_{\gamma,t}(y_{\gamma,t}(x)) + v_{\gamma,t}(y_{\gamma,t}(x))\partial_x F_{\gamma,t}(y_{\gamma,t}(x)) = 0$ for all $t \in [0,1]$, meaning that $F_{\gamma,t}(y_{\gamma,t}(x))$ is constant for all $t \in [0,1]$. This yields the identity:

$$\Phi\left(\frac{x}{\tau\sqrt{\gamma}}\right) = F_{\gamma,1}(y_{\gamma,1}(x)) = F_{\gamma,t}(y_{\gamma,t}(x)).$$

Thus,

$$\begin{aligned} y_{\gamma,t}(x) &= F_{\gamma,t}^{-1}\left(\Phi\left(\frac{x}{\tau\sqrt{\gamma}}\right)\right), \\ y_{\gamma,t}^{-1}(x) &= \tau\sqrt{\gamma}\Phi^{-1}(F_{\gamma,t}(x)). \end{aligned}$$

Since $\Sigma = \tau^2$, we have $\|y_{\gamma,t}^{-1}(x)\|_{\Sigma^{-1}}^2 = \gamma[\Phi^{-1}(F_{\gamma,t}(x))]^2$. Substituting this into Equation (3.6), we get the result. □

Definition C.2 (Symmetrically preserved ray)**.** Let $\nu_k := \Sigma^{-1/2}\mu_k$ be the component means in whitened coordinates, with weights π_k and $\Sigma \succ 0$. For a unit vector $\hat{e}$, write

$$e := \frac{\Sigma^{1/2}\hat{e}}{\|\Sigma^{1/2}\hat{e}\|}.$$

We call $\{re : r \geq 0\}$ a *symmetrically preserved ray* if there is a group H of orthogonal transformations preserving the weighted distribution of the ν_k, whose common fixed subspace is $\mathrm{span}\{\hat{e}\}$. Explicitly, require

$$\sum_k \pi_k \delta_{Q\nu_k} = \sum_k \pi_k \delta_{\nu_k} \qquad (Q \in H), \tag{C.7}$$

$$\mathrm{span}\{\hat{e}\} = \{z \in \mathbb{R}^n : Qz = z \text{ for every } Q \in H\}. \tag{C.8}$$

Theorem C.3 (Analysis along symmetrically preserved rays)**.** *Assume the setting of Theorem 3.4, with $\bar{\mu} = 0$, $R > 0$, $\gamma > 1$, $\sum_k \pi_k\mu_k = 0$ and $\Sigma \succ 0$. Let $\{re : r \geq 0\}$ be a symmetrically preserved ray as in Definition C.2, write $\lambda := e^\top\Sigma^{-1}e$ and define the scalar velocity $\hat{v}_{\gamma,t}(r) := \langle e, v_{\gamma,t}(re)\rangle$. Then,*

1. *The ray $\{re : r \geq 0\}$ is invariant under the guided flow $v_{\gamma,t}$ for all $\gamma \geq 0, t \in [0,1]$.*
2. *Let $q_{\gamma,t}$ be the unique probability density on $\mathbb{R}$ satisfying,*

$$\partial_t q_{\gamma,t} + \partial_r(\hat{v}_{\gamma,t}q_{\gamma,t}) = 0, \qquad q_{\gamma,1}(r) = \mathcal{N}(r; 0, \gamma/\lambda). \tag{C.9}$$

Define $F_{\gamma,t}(r) := \int_{-\infty}^{r} q_{\gamma,t}(u)\mathrm{d}u$. Then for every $t \in [0,1]$ and $r \in \mathbb{R}$ the weighting function $W_{\gamma,t}(x)$ for $p_{\gamma,t}(x|c)$ as defined in (3.8) satisfies,

$$W_{\gamma,t}(re) = \gamma^{n/2}\exp\left(-\frac{\gamma-1}{2}\left[\Phi^{-1}(F_{\gamma,t}(r))\right]^2\right). \tag{C.10}$$

Moreover $F_{\gamma,t,e}(0) = 1/2$ and $W_{\gamma,t}(re)$ is strictly decreasing in $r > 0$.

3. *Provided the orthogonal components of the whitened means have equal magnitudes, $q_{\gamma,t}$ can be written as a Gaussian mixture. Explicitly, let $a_k := \hat{e}^\top \nu_k$ and suppose that for some $B \geq 0$ and all k, $\|\nu_k - a_k \hat{e}\|^2 = B^2$. Then,*

$$q_{\gamma,t}(r) = \sum_k \pi_k \mathcal{N}\left(r; \frac{\gamma \alpha(t) a_k}{\sqrt{\lambda}}, \frac{\gamma}{\lambda}\right). \tag{C.11}$$

Proof. **Proof of 1.** In whitened coordinates $z = \Sigma^{-1/2} x$, the guided velocity is

$$\widetilde{v}_{\gamma,t}(z) = \gamma \dot{\alpha}(t) \frac{\sum_k \pi_k e^{\alpha(t) z^\top \nu_k} \nu_k}{\sum_k \pi_k e^{\alpha(t) z^\top \nu_k}}. \tag{C.12}$$

Let H be a symmetry group as in Definition C.2. The invariance in (C.7) implies $\widetilde{v}_{\gamma,t}(Qz) = Q\widetilde{v}_{\gamma,t}(z)$ for every $Q \in H$. Along $z = r\hat{e}$, applying both properties implies that

$$\widetilde{v}_{\gamma,t}(r\hat{e}) = \widetilde{v}_{\gamma,t}(Qr\hat{e}) = Q\widetilde{v}_{\gamma,t}(r\hat{e}).$$

Since this holds for every $Q \in H$, the common fixed-subspace assumption implies that $\widetilde{v}_{\gamma,t}(r\hat{e})$ lies in $\mathrm{span}\{\hat{e}\}$. Converting back to the original coordinates gives $v_{\gamma,t}(re) \in \mathrm{span}\{e\}$.

The restriction of the flow to this line is the flow of a scalar ODE. With $\hat{v}_{\gamma,s}(r) = \langle e, v_{\gamma,s}(re) \rangle$, fix t and define $\rho_{t,s}$ as the transport map arising from integrating the scalar velocity from t to s and $h_t(r) := \rho_{t,1}(r)$ as the transport from t to 1.

$$\frac{\mathrm{d}}{\mathrm{d}s}\rho_{t,s}(r) = \hat{v}_{\gamma,s}(\rho_{t,s}(r)), \qquad \rho_{t,t}(r) = r, \qquad h_t(r) := \rho_{t,1}(r). \tag{C.13}$$

The scalar velocity $\hat{v}_{\gamma,t}$ and its spatial derivative are bounded, so this ODE has a unique smooth flow given an initial condition. Since $v_{\gamma,t}(re) \in \mathrm{span}\{e\}$ we can write,

$$y_{\gamma,t}^{-1}(re) = h_t(r)e, \tag{C.14}$$

The mean-zero assumption implies $v_{\gamma,t}(0) = 0$ for all t and thus $h_t(0) = 0$. Since h_t is injective and $h_t'(0) > 0$ (as will be shown below), $h_t(r) > 0$ for $r > 0$, meaning that the positive ray $\{re\}$ is preserved.

Proof of 2. Having defined $h_t(r)$, we proceed analogously to Corollary 3.5. The density $q_{\gamma,t}$ in (C.9) is the transport of its Gaussian density by this scalar flow. Integrating the continuity equation for $q_{\gamma,t}$ from $-\infty$ to r (as in the proof of Corollary 3.5), yields,

$$\partial_t F_{\gamma,t} + \hat{v}_{\gamma,t}\, \partial_r F_{\gamma,t} = 0. \tag{C.15}$$

Observe that $\frac{\mathrm{d}}{\mathrm{d}s} F_{\gamma,s}(\rho_{t,s}(r)) = \partial_s F_{\gamma,s}(\rho_{t,s}(r)) + \hat{v}_{\gamma,s}\, \partial_r F_{\gamma,s}(\rho_{t,s}(r)) = 0$ which means that $F_{\gamma,s}(\rho_{t,s}(r))$ is constant for all s. Using $s = t$ (for which $\rho_{s,s}$ is identity) and $s = 1$ (for which $F_{\gamma,1}$ is the CDF of $\mathcal{N}(0, \gamma/\lambda)$), we get,

$$F_{\gamma,t}(r) = F_{\gamma,1}(\rho_{t,1}(r)) = \Phi\left(\sqrt{\lambda/\gamma}\, h_t(r)\right). \tag{C.16}$$

In particular, since $h_t(0) = 0$, $F_{\gamma,t}(0) = 1/2$. By (3.6) and (C.14),

$$W_{\gamma,t}(re) = \gamma^{n/2} \exp\left(-\frac{\gamma - 1}{2\gamma} \lambda h_t(r)^2\right). \tag{C.17}$$

Inverting (C.16) and substituting into (C.17) proves (C.10),

$$W_{\gamma,t}(re) = \gamma^{n/2} \exp\left(-\frac{\gamma - 1}{2} \left[\Phi^{-1}(F_{\gamma,t}(r))\right]^2\right).$$

Since the density of $q_{\gamma,1}$ is strictly positive and $v_{\gamma,t}$ is continuous, $q_{\gamma,t}$ also has strictly positive density for all t and the associated CDF, $F_{\gamma,t}$, is strictly increasing. As both Φ^{-1} and $F_{\gamma,t}$ are strictly increasing, $W_{\gamma,t}(re)$ must be strictly increasing for $r < 0$ and decreasing for $r > 0$. Note that this also implies that $h_t'(0) > 0$, completing the requirement we deferred from the proof of the first condition.

Proof of 3. Finally, suppose that $B \geq 0$ and all k, $\|\nu_k - a_k \hat{e}\|^2 = B^2$ holds. Since $\Sigma^{-1/2} e = \sqrt{\lambda}\, \hat{e}$, writing out the projected scalar velocity, the weights for the orthogonal components are the same for all k and therefore we obtain,

$$\hat{v}_{\gamma,t}(r) = \dot{\alpha}(t) \frac{\sum_k \pi_k e^{\alpha(t)\sqrt{\lambda}\, r a_k} \left[\frac{\gamma a_k}{\sqrt{\lambda}}\right]}{\sum_k \pi_k e^{\alpha(t)\sqrt{\lambda}\, r a_k}}. \tag{C.18}$$

We can directly observe from Proposition 3.3 that this is the flow for the mixture:

$$q_{\gamma,t}(r) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(r, \frac{\gamma \alpha(t) a_k}{\sqrt{\lambda}}, \frac{\gamma}{\lambda}\right).$$

□

Theorem 3.7 (Central Enhancement and Tail Suppression)**.** *Assume the setting of Theorem 3.4 with $n \geq 1$ and $\gamma > 1$. Then, for every $t \in [0,1]$ and any $x \in \mathbb{R}^n$, the density factor $W_{\gamma,t}(x)$ satisfies,*

$$\exp\left(-\frac{\gamma-1}{2\gamma}\|x - y_{\gamma,t}(0)\|^2_{\Sigma^{-1}}\right) \leq \gamma^{-n/2} W_{\gamma,t}(x) \leq \exp\left(-\frac{\gamma-1}{2\gamma} e^{-\gamma R^2 \alpha(t)^2} \|x - y_{\gamma,t}(0)\|^2_{\Sigma^{-1}}\right).$$

Note that at $x = y_{\gamma,t}(0)$ these bounds match and recover that $W_{\gamma,t}(y_{\gamma,t}(0)) = \gamma^{n/2}$.

Proof. For simplicity we can additionally assume $\bar{\mu} = 0$, as we can simply consider the centered coordinate space $x - \alpha(t)\bar{\mu}$ in which the mixture components lie on an ellipse centered at 0 with $\|\cdot\|_{\Sigma^{-1}}$ radius R and where guidance is equivalent to scaling the conditional flow.

High-level strategy: For this proof, we will show that,

$$e^{-\gamma R^2 \alpha(t)^2/2} \|x - y_{\gamma,t}(0)\|_{\Sigma^{-1}} \leq \|y^{-1}_{\gamma,t}(x)\|_{\Sigma^{-1}} \leq \|x - y_{\gamma,t}(0)\|_{\Sigma^{-1}},$$

holds for all x and $t \in [0,1]$. To construct these bounds, we use that $y^{-1}_{\gamma,t}(y_{\gamma,t}(0)) = 0$ and then bound $y^{-1}_{\gamma,t}(x)$ via sensitivity analysis of its spatial derivative, combined with integration along the segment from $y_{\gamma,t}(0)$ to x for the upper bound and from 0 to $y^{-1}_{\gamma,t}(x)$ for the lower bound.

Let D denote differentiation with respect to the spatial argument. Choose any z and tangent u and set $h_s := D y_{\gamma,s}(z) u$. Write $\langle a, b\rangle_{\Sigma^{-1}} := a^\top \Sigma^{-1} b$ and $m_s(z) := \sum_k w_k(s,z)\mu_k$. Differentiating the flow equation with respect to z gives $\dot{h}_s = D v_{\gamma,s}(y_{\gamma,s}(z)) h_s$ and $h_1 = u$. Consequently,

$$\frac{\mathrm{d}}{\mathrm{d}s}\|h_s\|^2_{\Sigma^{-1}} = 2\langle h_s, D v_{\gamma,s}(y_{\gamma,s}(z)) h_s\rangle_{\Sigma^{-1}}.$$

Differentiating the weights in (3.2) in any direction h gives

$$Dw_k(s,z)h = \alpha(s) w_k(s,z)\langle \mu_k - m_s(z), h\rangle_{\Sigma^{-1}},$$

$$\langle h, D v_{\gamma,s}(z) h\rangle_{\Sigma^{-1}} = \gamma \dot{\alpha}(s)\alpha(s)\left[\sum_k w_k(s,z)\langle \mu_k, h\rangle^2_{\Sigma^{-1}} - \langle m_s(z), h\rangle^2_{\Sigma^{-1}}\right].$$

Dropping the negative term and applying Cauchy–Schwarz yields

$$\sum_k w_k(s,z)\langle \mu_k, h\rangle^2_{\Sigma^{-1}} \leq \sum_k w_k(s,z)\|\mu_k\|^2_{\Sigma^{-1}}\|h\|^2_{\Sigma^{-1}} = R^2\|h\|^2_{\Sigma^{-1}}.$$

Since

$$\sum_k w_k(s,z)\langle \mu_k, h\rangle^2_{\Sigma^{-1}} - \langle m_s(z), h\rangle^2_{\Sigma^{-1}} = \sum_k w_k(s,z)\langle \mu_k - m_s(z), h\rangle^2_{\Sigma^{-1}} \geq 0,$$

and $\dot{\alpha}(s) \leq 0, \alpha(s) \geq 0$, this implies,

$$-\gamma|\dot{\alpha}(s)|\alpha(s) R^2 \|h\|^2_{\Sigma^{-1}} \leq \langle h, D v_{\gamma,s}(z) h\rangle_{\Sigma^{-1}} \leq 0.$$

Therefore,

$$-2\gamma|\dot{\alpha}(s)|\alpha(s) R^2 \|h_s\|^2_{\Sigma^{-1}} \leq \frac{\mathrm{d}}{\mathrm{d}s}\|h_s\|^2_{\Sigma^{-1}} \leq 0.$$

Note that,

$$\frac{\mathrm{d}}{\mathrm{d}s}\left[e^{2\int_t^s \gamma|\dot\alpha(r)|\alpha(r)R^2\mathrm{d}r}\|h_s\|_{\Sigma^{-1}}^2\right] = e^{2\int_t^s \gamma|\dot\alpha(r)|\alpha(r)R^2\mathrm{d}r}\left[2\gamma|\dot\alpha(s)|\alpha(s)R^2\|h_s\|_{\Sigma^{-1}}^2 + \frac{\mathrm{d}}{\mathrm{d}s}\|h_s\|_{\Sigma^{-1}}^2\right].$$

Thus the right inequality implies that $\|h_s\|_{\Sigma^{-1}}^2$ is non-increasing, while the left implies that $e^{2\int_t^s \gamma|\dot\alpha(r)|\alpha(r)R^2\mathrm{d}r}\|h_s\|_{\Sigma^{-1}}^2$ is non-decreasing. Since $\dot\alpha \le 0$ and $\alpha(1) = 0$, let $L_t := \int_t^1 \gamma|\dot\alpha(r)|\alpha(r)R^2\mathrm{d}r = \gamma R^2\alpha(t)^2/2$. Comparing their values at t and 1 yields,

$$e^{-2L_t}\|h_t\|_{\Sigma^{-1}}^2 \le \|h_1\|_{\Sigma^{-1}}^2 \le \|h_t\|_{\Sigma^{-1}}^2,$$

Taking the square root, rearranging, and using that $h_t = Dy_{\gamma,t}(z)u$, $h_1 = u$ gives the inequality:

$$\|u\|_{\Sigma^{-1}} \le \|Dy_{\gamma,t}(z)u\|_{\Sigma^{-1}} \le e^{L_t}\|u\|_{\Sigma^{-1}}. \tag{C.19}$$

Converting the bounds to the inverse mapping. To complete our argument, ultimately we want bounds on $\|Dy_{\gamma,t}^{-1}(z)u\|_{\Sigma^{-1}}$ instead of $\|Dy_{\gamma,t}(x)u\|_{\Sigma^{-1}}$. Differentiating $y_{\gamma,t}(y_{\gamma,t}^{-1}(x)) = x$ by the chain rule gives that for all $\tilde u$,

$$Dy_{\gamma,t}(y_{\gamma,t}^{-1}(x))Dy_{\gamma,t}^{-1}(x)\tilde u = \tilde u.$$

Applying the lower derivative bound of (C.19) at the point $z = y_{\gamma,t}^{-1}(x)$ and in the direction $u = Dy_{\gamma,t}^{-1}(x)\tilde u$ gives

$$\|Dy_{\gamma,t}^{-1}(x)\tilde u\|_{\Sigma^{-1}} \le \|Dy_{\gamma,t}(y_{\gamma,t}^{-1}(x))Dy_{\gamma,t}^{-1}(x)\tilde u\|_{\Sigma^{-1}} = \|\tilde u\|_{\Sigma^{-1}}.$$

For all t, we have $y_{\gamma,t}^{-1}(y_{\gamma,t}(0)) = 0$. Therefore we can bound via integration along the segment $\{\lambda x + (1-\lambda)y_{\gamma,t}(0) : \lambda \in [0,1]\}$ from $y_{\gamma,t}(0)$ to x.

$$\begin{aligned}
y_{\gamma,t}^{-1}(x) &= y_{\gamma,t}^{-1}(x) - y_{\gamma,t}^{-1}(y_{\gamma,t}(0)) \\
&= \int_0^1 Dy_{\gamma,t}^{-1}(\lambda x + (1-\lambda)y_{\gamma,t}(0))[x - y_{\gamma,t}(0)]\,\mathrm{d}\lambda, \\
\|y_{\gamma,t}^{-1}(x)\|_{\Sigma^{-1}} &\le \int_0^1 \|Dy_{\gamma,t}^{-1}(\lambda x + (1-\lambda)y_{\gamma,t}(0))[x - y_{\gamma,t}(0)]\|_{\Sigma^{-1}}\,\mathrm{d}\lambda \\
&\le \int_0^1 \|x - y_{\gamma,t}(0)\|_{\Sigma^{-1}}\,\mathrm{d}\lambda \\
&= \|x - y_{\gamma,t}(0)\|_{\Sigma^{-1}}.
\end{aligned}$$

For the other inequality, apply the same argument to $y_{\gamma,t}$ along the segment from 0 to $y_{\gamma,t}^{-1}(x)$, using its upper derivative bound:

$$\begin{aligned}
x - y_{\gamma,t}(0) &= y_{\gamma,t}(y_{\gamma,t}^{-1}(x)) - y_{\gamma,t}(y_{\gamma,t}^{-1}(y_{\gamma,t}(0))) \\
&= y_{\gamma,t}(y_{\gamma,t}^{-1}(x)) - y_{\gamma,t}(0) \\
&= \int_0^1 Dy_{\gamma,t}(\lambda y_{\gamma,t}^{-1}(x))y_{\gamma,t}^{-1}(x)\,\mathrm{d}\lambda, \\
\|x - y_{\gamma,t}(0)\|_{\Sigma^{-1}} &\le \int_0^1 \|Dy_{\gamma,t}(\lambda y_{\gamma,t}^{-1}(x))y_{\gamma,t}^{-1}(x)\|_{\Sigma^{-1}}\,\mathrm{d}\lambda \\
&\le \int_0^1 e^{L_t}\|y_{\gamma,t}^{-1}(x)\|_{\Sigma^{-1}}\,\mathrm{d}\lambda = e^{L_t}\|y_{\gamma,t}^{-1}(x)\|_{\Sigma^{-1}}.
\end{aligned}$$

Dividing the last inequality by e^{L_t} and combining the two bounds yields,

$$e^{-\gamma R^2\alpha(t)^2/2}\|x - y_{\gamma,t}(0)\|_{\Sigma^{-1}} \le \|y_{\gamma,t}^{-1}(x)\|_{\Sigma^{-1}} \le \|x - y_{\gamma,t}(0)\|_{\Sigma^{-1}}.$$

Squaring these bounds and substituting into (3.6) therefore yields,

$$\gamma^{n/2}\exp\left(-\frac{\gamma-1}{2\gamma}\|x - y_{\gamma,t}(0)\|_{\Sigma^{-1}}^2\right) \le W_{\gamma,t}(x) \le \gamma^{n/2}\exp\left(-\frac{\gamma-1}{2\gamma}e^{-\gamma R^2\alpha(t)^2}\|x - y_{\gamma,t}(0)\|_{\Sigma^{-1}}^2\right).$$

□

D EXPERIMENT DETAILS

Models and Sampling. We use pretrained Stable Diffusion 1.5 and FLUX.2-klein-base-9B for text-conditioned generation. The ImageNet (SD-VAE) experiments use a pretrained DiT-XL/2 model. For the ImageNet (FLUX.2-VAE) experiments, we trained a DiT-B/1 model ourselves and use its EMA weights for sampling. We generate text-conditioned images at 512×512 and ImageNet images at 256×256.

Each image experiment uses 10,000 samples per condition and guidance scale, with the same initial noise across scales. The density comparisons use $\gamma \in \{0, 0.5, 1, 1.5, 2, 4, 8, 12\}$, with additional values 6.5 and 10 for the ImageNet (SD-VAE) experiments. We use four ImageNet classes: golden retriever, Arctic fox, fire engine, and strawberry. The text prompts are "a studio photograph of a red fox in a snowy forest," "a watercolor painting of a sailboat at sunset," "a futuristic city street at night," and "a ceramic teapot on a wooden table."

Noise Schedules. For all models we use their "native" noise schedules as used by upstream repositories and configuration. The SD1.5 and ImageNet (SD-VAE) experiments use variance-preserving schedules with $\alpha(t_j) = \sqrt{\prod_{i=1}^{j}(1-\beta_i)}$ and $\sigma(t_j) = \sqrt{1-\alpha(t_j)^2}$ on 1,000 training noise levels. For SD1.5, $\sqrt{\beta_i}$ is linearly spaced between $\sqrt{0.00085}$ and $\sqrt{0.012}$; for ImageNet (SD-VAE), β_i is linearly spaced from 0.0001 to 0.02. Both use 50-step deterministic DDIM sampling. The FLUX Klein and ImageNet with FLUX.2 latents experiments use linear flow matching, $\alpha(t) = 1 - t$ and $\sigma(t) = t$. FLUX Klein uses 20 Euler steps with the model's default exponential time shift. The ImageNet with FLUX.2 latents experiments use 50 Euler steps at uniformly spaced times from 1 to 0.18, followed by 0.

Cluster Projections. We use half the samples for fitting and half for evaluation. For each condition, we reduce flattened $\gamma = 1$ latents to 32 principal components and fit k-means, selecting 1–16 clusters by spherical-Gaussian BIC. Assignments remain fixed across guidance, and projection directions use the fitting half. We retain clusters with at least 24 samples in each half. The principal direction in Figure 5 points from the unconditional mean at $\gamma = 0$ to the cluster mean at $\gamma = 12$. The control direction is sampled independently at each scale, orthogonal to this displacement. For Figure 6, we instead project from the condition mean toward the cluster mean at each scale. We standardize the held-out projections within each cluster and pool them, giving equal weight to conditions and to clusters within each condition. We estimate densities with Gaussian KDE at bandwidth 0.15, correcting for the variance added by smoothing. The principal plots restore the weighted cluster mean and within-cluster variance, expressed relative to $\gamma = 1$. The radial plots retain zero mean and unit variance, with skewness computed as the pooled third moment.

t-SNE Embeddings. For Figures 2 and 8, we visualize two-dimensional t-SNE embeddings of DINO-V2 features extracted from generated ImageNet images, with 10 different classes in each plot. We use a diffusion model pretrained with the Flux 2 VAE, with samples generated using DDIM without classifier-free guidance. Conditional samples are generated using class conditioning. Unconditional samples are generated by conditioning on the "null" class and selected by rejection sampling with a ResNet-based classifier. The t-SNE plots are each generated in an embedding space using the conditional and unconditional samples corresponding to the 10 selected classes as well as a set of unconditional samples outside of the selected classes. The plots in Figure 8 repeat this comparison across different sets of classes.

Synthetic Experiments. We use the variance-preserving schedule $\alpha(t) = 1 - t$, $\sigma(t) = \sqrt{1-(1-t)^2}$ and integrate the analytic Gaussian-mixture vector fields from $t = 1$ to $t = 0$ using fourth-order Runge–Kutta with uniform steps. In Figure 1, the standard setup conditions an unconditional mixture on one component; our setup uses a Gaussian unconditional distribution and a mixture conditional distribution. The one-dimensional examples use conditional $\frac{1}{2}\mathcal{N}(-3, 1) + \frac{1}{2}\mathcal{N}(3, 1)$ and unconditional $\mathcal{N}(0, 1)$ distributions, with 256 integration steps. Figure 4 compares exact densities with histograms of 100,000 held-out samples per scale. The synthetic radial panel uses 2.5 million held-out samples per scale. We assign each sample to its nearest component at $\gamma = 1$ and project along that component's radial direction, then standardize the projections and plot smoothed histograms.

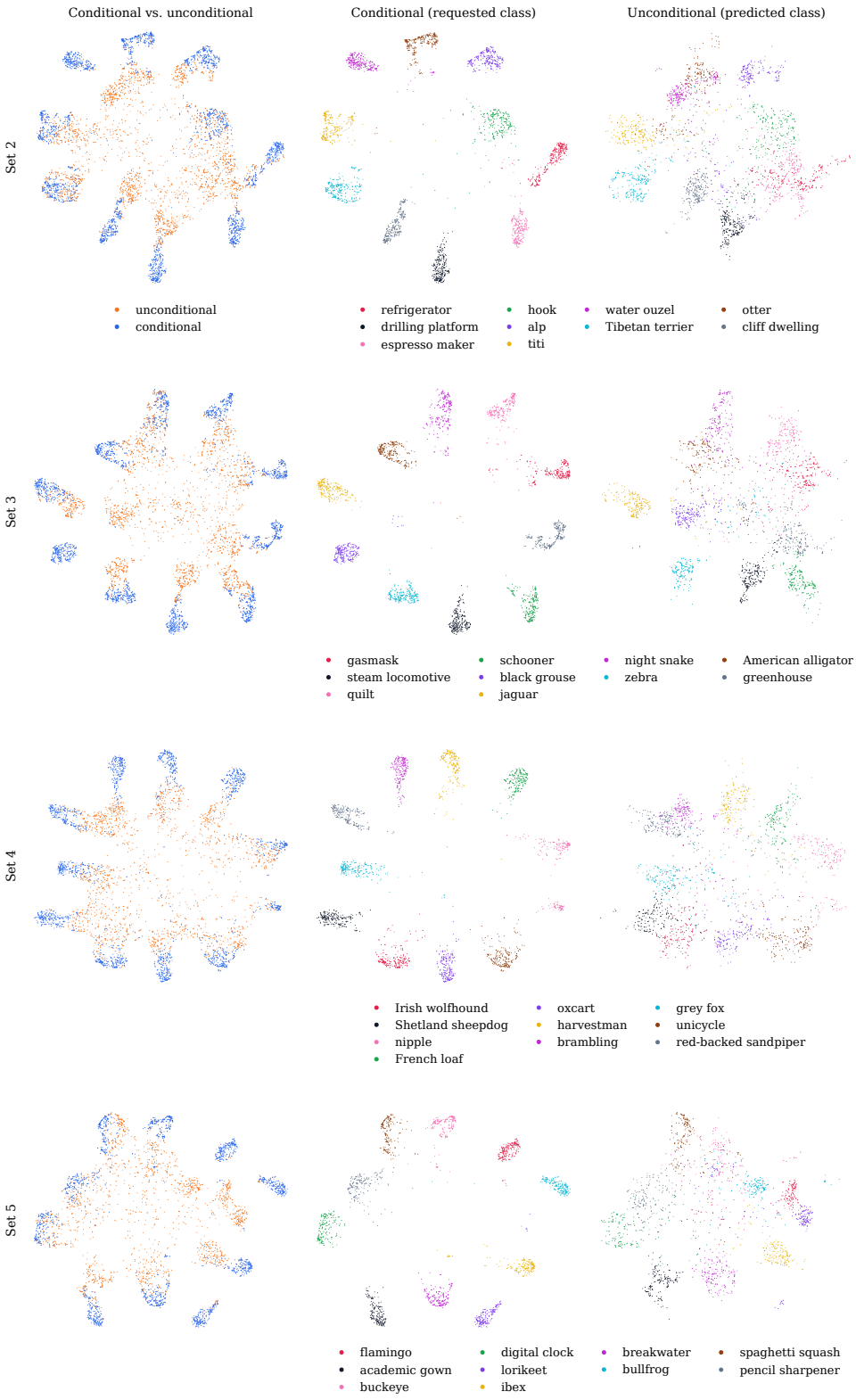

Figure 8: Scatter plots of t-SNE DINO-V2 embeddings for ImageNet conditional and unconditional generations, with 10 different classes in each plot. This shows that the multimodality and complex structure of conditional embeddings relative to unconditional embeddings is robust across different classes.

E SAMPLE GALLERIES

Sample Galleries. We construct mean images by averaging all 10,000 final latents for each condition and scale, then decoding the result. Zero-noise images use the same sampler initialized at $x_1 = 0$. The galleries show randomly selected samples at $\gamma = 1, 4, 8, 12$, with initial noise matched across columns.

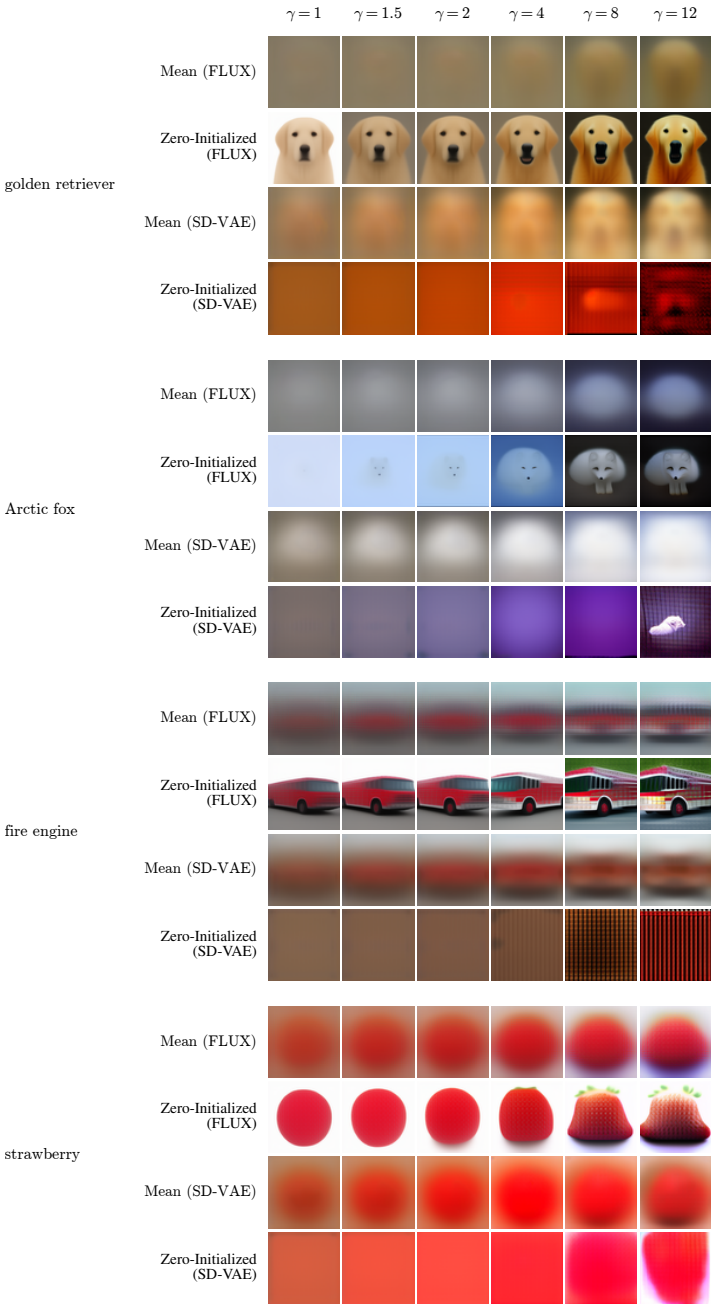

Figure 9: Mean sample and zero-initialized ($y_{\gamma,0}(0)$) comparisons for ImageNet classes: golden retriever, Arctic fox, fire engine, and strawberry. Columns show guidance scales γ; rows show FLUX and SD-VAE results, as labeled. While the zero-noise samples track the mean samples for FLUX, for imagenet, the zero-initialized samples are highly unstable.

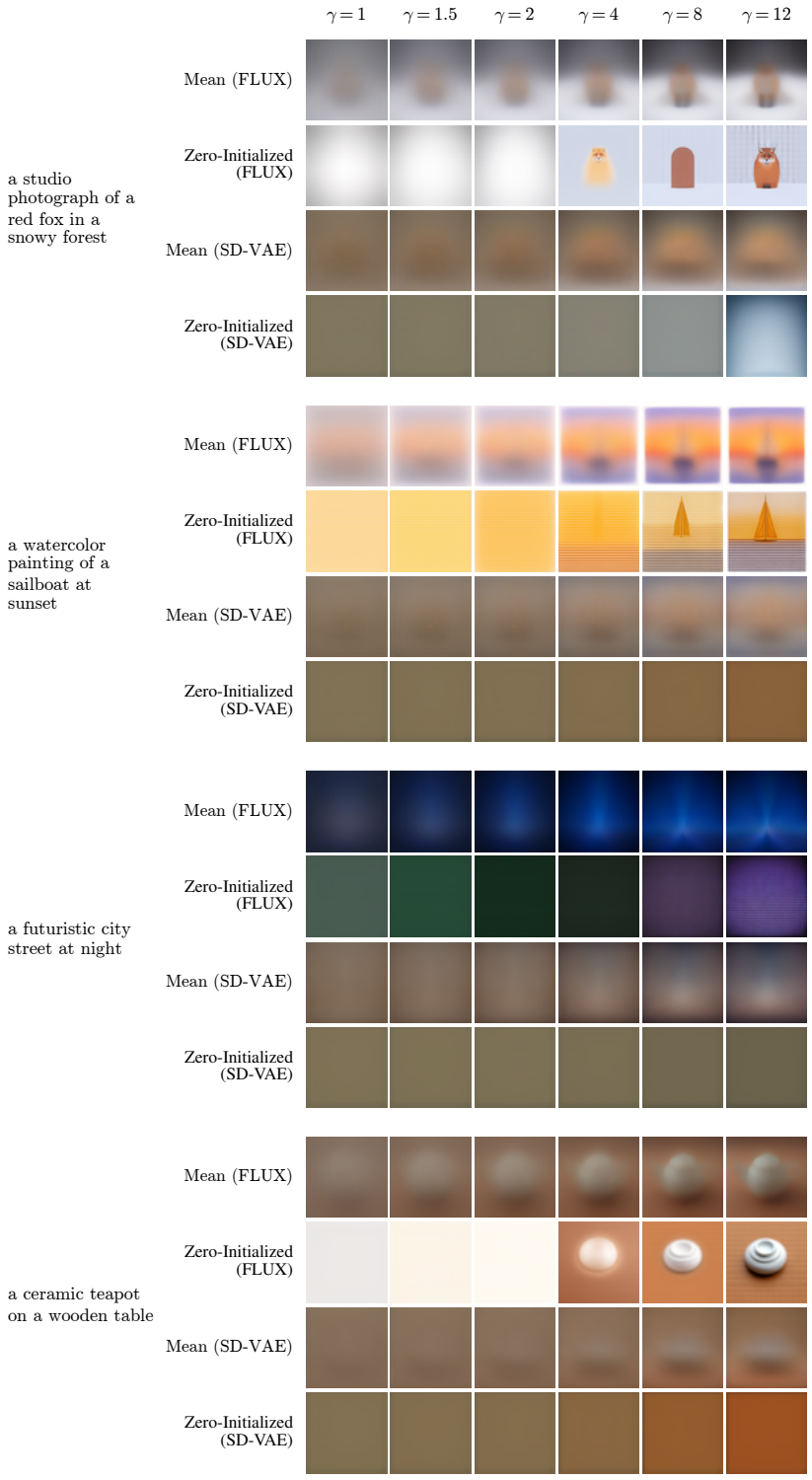

Figure 10: Mean sample and zero-initialized samples for the four text prompts shown at left. Columns show guidance scales γ; rows show FLUX and SD-VAE results, as labeled.

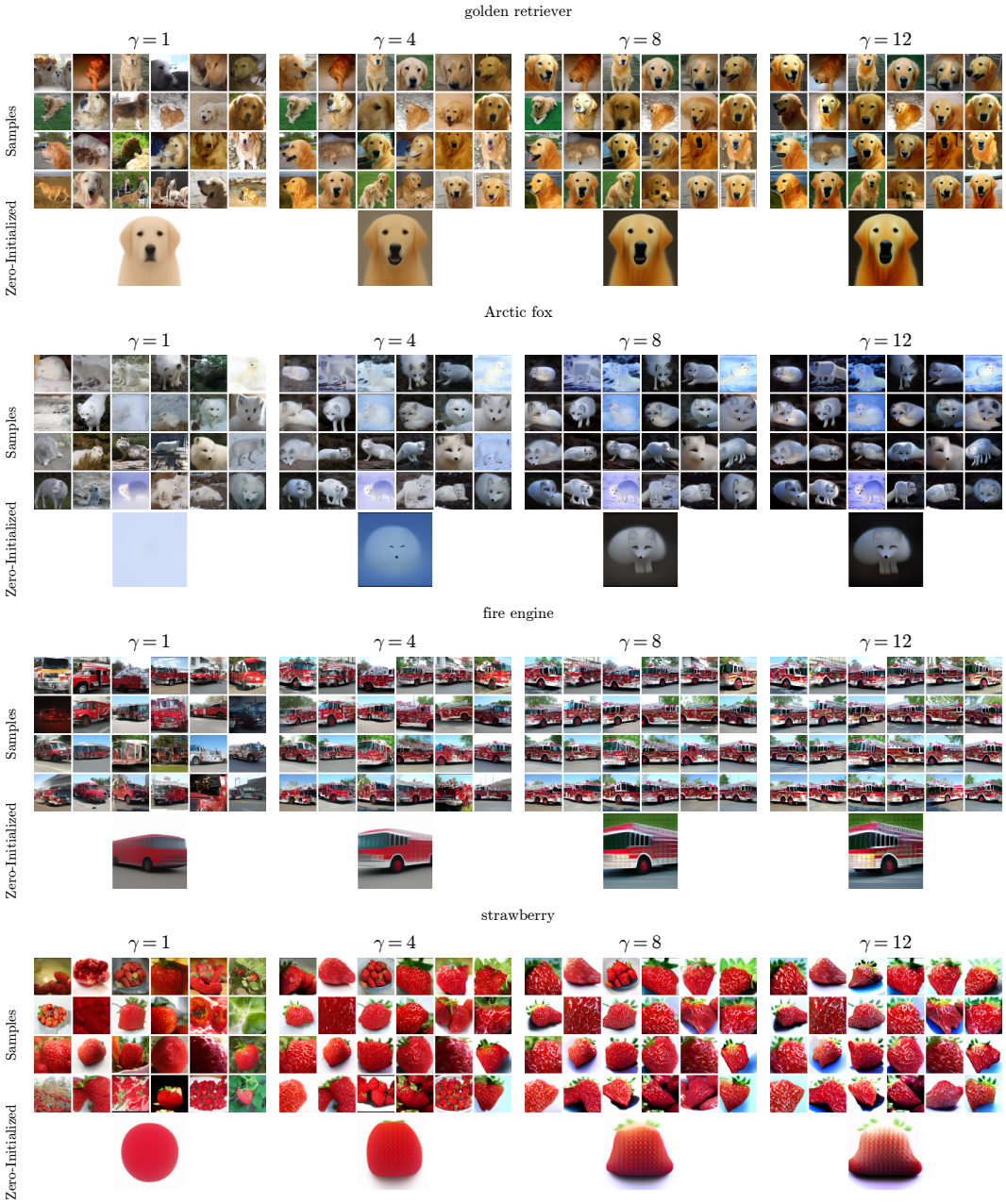

Figure 11: ImageNet FLUX generated samples and zero-noise outputs at guidance scales $\gamma = 1, 4, 8, 12$ for golden retriever, Arctic fox, fire engine, and strawberry. Rows are labeled Samples and Zero Noise.

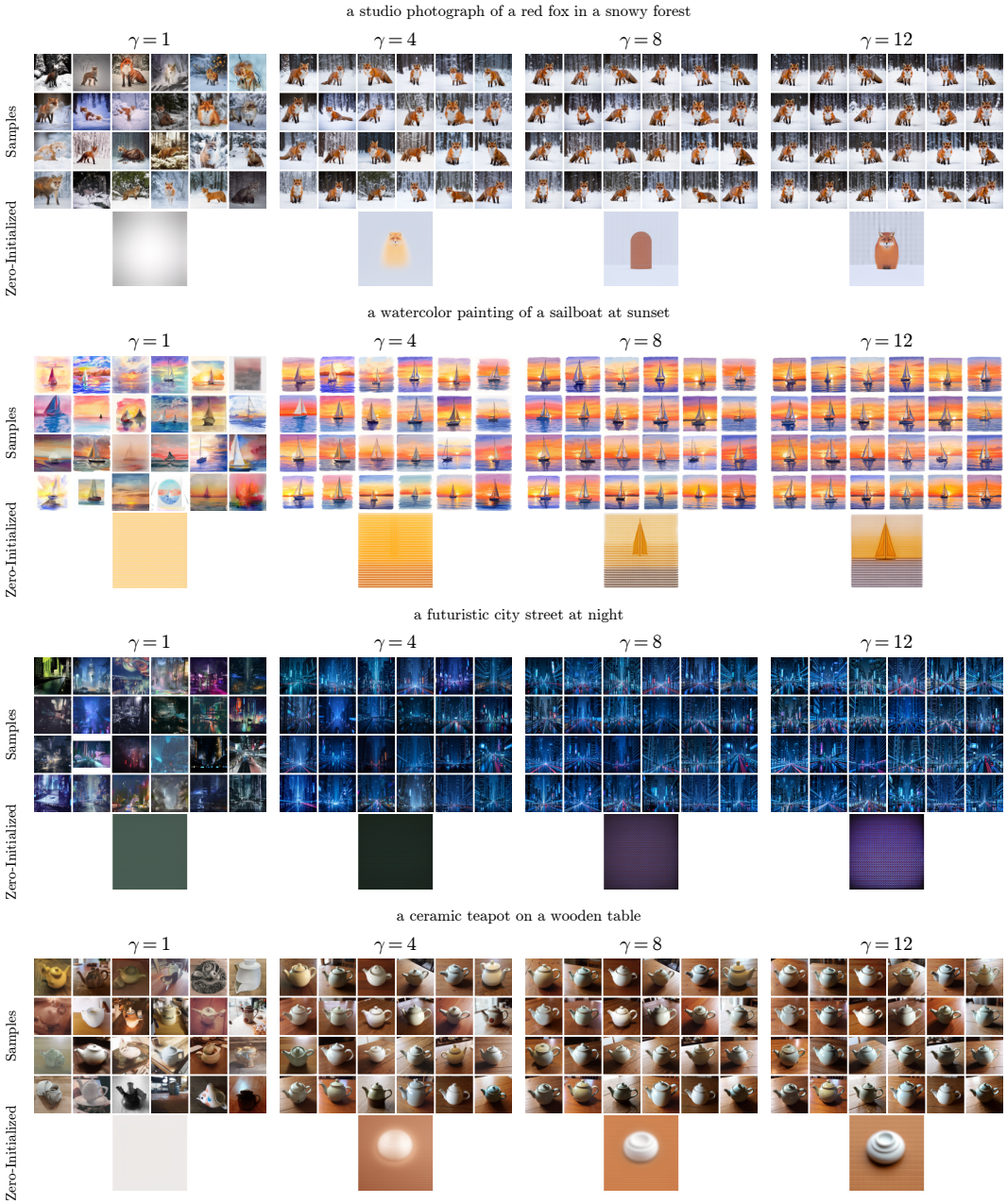

Figure 12: FLUX Klein generated samples and zero-noise outputs at guidance scales $\gamma = 1, 4, 8, 12$ for the four text prompts indicated above each panel.

F LANGUAGE MODEL USAGE

Large language models (LLMs) were used for manuscript and proof review, as well as for light editing of the main body and documenting hyperparameters in the experimental details section. Proofs in this manuscript were written and checked by hand, although LLMs were used to check for inconsistencies and issues. LLM-enabled search tools were used for finding some related work, although all cited work was at least in part read manually.

Language models and agentic tooling were heavily used in writing code for the experiments and to create the figures presented in this manuscript. We include all code necessary for reconstructing these figures in the supplementary information.