UNSUPERVISED LEARNING OF DISENTANGLED REPRESENTATIONS VIA DIFFUSION VARIATIONAL AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We present the diffusion variational autoencoder (DiVA), a method for unsupervised representation learning that combines the theoretical frameworks of diffusion models and VAEs. By unifying their respective evidence lower bounds, DiVA formulates a principled objective that learns representations through score-based guidance of the underlying diffusion process. The resulting representations automatically capture meaningful structure in the data: it recovers ground truth generative factors in synthetic datasets, learns factorized, semantic latent dimensions from complex natural images, and encodes video sequences into latent trajectories that are straighter than those of alternative encoders, despite training exclusively on static images. Furthermore, DiVA can extract useful representations from pretrained diffusion models with minimal additional training. Finally, the explicitly probabilistic formulation provides new ways to identify semantically meaningful axes in the absence of supervised labels. Overall, these results indicate that implicit structural information in diffusion models can be made explicit and interpretable through synergistic combination with a variational autoencoder.

1 INTRODUCTION

To evaluate the behavioral relevance of their observations intelligent agents must possess a notion of semantics. This requires decomposing complex sensory observations into abstract and semantically meaningful factors of variation, a capability known in machine learning as disentangling (Bengio et al., 2014). Recent theoretical and empirical work has demonstrated that disentangled representations offer several advantages, including improved interpretability, enhanced generalization, and stronger transfer capabilities across related tasks (Higgins et al., 2017). However, given that for both biological and artificial agents, the amount of unlabeled data vastly exceeds that of labeled data, an important consideration is how to learn disentangled representations without relying on explicit supervision.

Latent variational models provide a principled framework grounded in probabilistic generative models. The variational auto-encoder (VAE) has emerged as a particularly promising instantiation of this idea (Kingma and Welling, 2013; Rezende et al., 2014), which allows for amortized inference of latent structure and data generation through optimization of the evidence lower bound (ELBO). Successful variants of this approach have been used for compression (Ballé et al., 2017; Ballé et al., 2021), prediction (Tishby et al., 2000; Alemi et al., 2019; Sachdeva et al., 2020), and as models of neural activity in visual cortex (Csikor et al., 2022; Vafaii et al., 2023). Crucially, VAEs can be encouraged to exhibit disentangled representations via appropriate regularization, as with β-VAEs (Higgins et al., 2017; Alemi et al., 2019).

Despite these successes, there is a fundamental tension between reconstruction fidelity and disentanglement quality (Kumar et al., 2018; Sikka et al., 2019) in such models. Encouraging disentangled representations typically requires stronger regularization that limits the expressiveness of latent variables; this trade-off between compression and reconstruction precision is well known as the rate-distortion principle. This limitation has motivated various VAE improvements, including enhanced posterior approximations through non-diagonal covariances or complex posterior distributions (Manduchi et al., 2023; Klushyn et al., 2019; Mathieu et al., 2019; Cheng et al., 2020; Rezende and Mohamed, 2015), additional latent constraints (Chen et al., 2019; Zhao et al., 2018), and more

flexible latent priors (Klushyn et al., 2019; Wehenkel and Louppe, 2021). Notably, these efforts have largely focused on the inference mechanism rather than the generative model itself, despite the role the diagonal posterior covariance plays in encouraging disentanglement (Rolínek et al., 2019) and the importance of the generative model in determining the nature of learned representations (Cremer et al., 2018).

A possible avenue for building more expressive generative models lies in diffusion models. These generative models produce high-quality samples in many domains by learning to approximate the data distribution through a denoising objective. However, their lack of an explicit latent representation poses challenges for tasks requiring embeddings, such as disentangled representation learning (Song et al., 2020). Additionally, guiding the diffusion sampling process to generate samples with specific attributes or classes remains an open research question (Fuest et al., 2024). While techniques like classifier-guided diffusion (Dhariwal and Nichol, 2021) and classifier-free guidance (Ho and Salimans, 2021) lay the groundwork for sampling from conditional densities in the presence of explicit class labels, achieving robust and flexible conditioning without compromising sample quality or diversity is not fully resolved (Chidambaram et al., 2024; Sadat et al., 2024; Kaiser et al., 2024; Ifriqi et al., 2025). Critically, few studies explore the learning of latents that enable diffusion models to effectively navigate the full data manifold *without auxiliary information*.

Here, we introduce a Diffusion Variational Autoencoder (DiVA) that combines the strengths of variational autoencoders and diffusion models to achieve unsupervised learning of structured, interpretable latent representations while maintaining high-quality sample generation. Our key innovation lies in employing conditional diffusion as the generative component of a VAE, coupled with a precise mathematical formalization of the objective via the ELBO that enables reuse of the inference network for conditioning. The diffusion prior on the data space allows us to keep a simple latent posterior approximation and impose strong factorization constraints on the latent code. Furthermore, learning latent representations of images that are coherent across noise levels helps to unify and smooth the latent representation. When trained on image data, our framework yields semantically meaningful latent axes and produces straighter trajectories when encoding natural video sequences, all without compromising generative performance. Moreover, a distinctive feature of our approach is that the posterior structure enables identification of semantically meaningful axes without reliance on supervised labels. From the diffusion perspective, incorporating an inference network with an explicit representational space facilitates unsupervised discovery of factorized and interpretable latents that can be used to control sample generation. Notably, this same methodology can be applied to extract semantic latent representations from pretrained diffusion models.

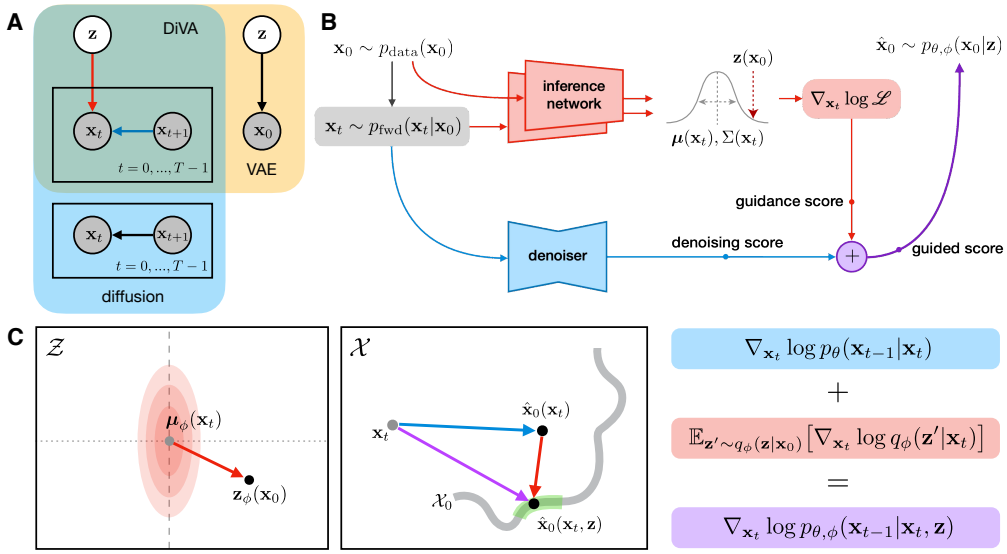

Figure 1: **A)** Graphical model for our diffusion VAE (DiVA), contrasted with those of standard VAEs and unconditioned diffusion models. **B)** Schematic of conditional sampling procedure. **C)** Schematic illustration of how movement in the latent space results in guidance of the denoiser score towards regions of the clean image manifold that are semantically similar to the original image. For visualization, both latent and image spaces are depicted as two dimensional.

2 METHODS

The task of extracting latent representations from data can be formalized as a graphical model in which latent random variables, $\mathbf{z} \in Z$, give rise to observations $\mathbf{x} \in X$ (Fig. 1A, yellow box). Representational learning then corresponds to maximizing the marginal likelihood of the observed data under the model, $\prod_k \int p_\theta(\mathbf{x}_k, \mathbf{z}) d\mathbf{z}$, while inference requires calculating the posterior distribution $p(\mathbf{z}|\mathbf{x})$ via Bayes' rule. Unfortunately, directly computing the log likelihood and the posterior distribution in closed form is usually infeasible for expressive models.

Variational autoencoders (VAEs) address this difficulty by approximating the posterior with a simpler parametric form, whose parameters are data-specific and obtained via optimization (Kingma and Welling, 2013; Rezende et al., 2014). The parameters of both the amortized posterior and the generative model are learned via minimization of the *evidence lower bound* (ELBO), a tractable lower bound on the model marginal likelihood:

$$\mathcal{L}_{\text{ELBO}}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right] \leq \log p_\theta(\mathbf{x}), \tag{1}$$

where $q_\phi(\mathbf{z}|\mathbf{x})$ is the approximate posterior (usually computed by a neural network with weights ϕ), the density $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})$ is the generative model, with prior $p(\mathbf{z})$ given by a simple parametric distribution (e.g., a standard Gaussian), and likelihood $p_\theta(\mathbf{x}|\mathbf{z})$ computed by a separate neural network with weights θ. In practice, the limitations imposed on the likelihood by this parameterization often lead to poor quality of generated samples (Burda et al., 2016; Sønderby et al., 2016; Rezende and Viola, 2018).

In contrast, diffusion models are highly expressive generative models that can estimate and sample from complex densities. Diffusion models such as denoising diffusion probabilistic models (DDPMs) form an implicit prior over data by learning to denoise noisy versions of clean data (Sohl-Dickstein et al., 2015; Kadkhodaie and Simoncelli, 2020; Ho et al., 2020). Noisy data are generated by a fixed *forward* operator that is typically assumed to be additive Gaussian, $\mathbf{x}_t \sim p_{\text{fwd}}(\mathbf{x}_t|\mathbf{x}_0)$, with $t \in [0, 1, ..., T]$ specifying the noise variance. This denoising objective can be written as the mean squared error between the model's estimate of the noise $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ and the true noise present in the noisy image $\boldsymbol{\epsilon}$, which is typically assumed to be isotropic Gaussian. Minimizing this error has also been shown to be equivalent to maximizing the ELBO (up to a noise-dependent scalar weighting λ_t) for a joint data distribution given by $p(\mathbf{x}_{0:T})$ and a joint "posterior" $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ that captures the distribution over all noising paths (Kingma and Gao, 2023):

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(0,I),\, t \sim [0,T]} \left[\lambda_t \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2 \right] = -\mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right]. \tag{2}$$

It is well known that the minimum mean squared error (MMSE) estimate is given by the posterior mean, and this in turn can be related to the score function via Miyasawa's/Tweedie's formula (Robbins, 1956; Miyasawa et al., 1961). Under a Gaussian variance-preserving noise distribution given by $p(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t|\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)I)$, where $\{\bar{\alpha}_t\}$ defines the variance schedule, the MMSE estimate of the noise is given as

$$\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t) = \underset{\boldsymbol{\epsilon}_\theta}{\operatorname{argmin}} \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2 = \mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_t)}[\boldsymbol{\epsilon}] = -\sqrt{1-\bar{\alpha}_t} \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t). \tag{3}$$

Once trained, generation of samples relies on an iterative partial denoising process that samples $\mathbf{x}_t \sim p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})$ as $t = T-1, ..., 0$ (Fig. 1A, blue box). Examples from the training set can be thought of as samples from a complex probability distribution that resembles a low-dimensional, nonlinear manifold lying within the high-dimensional data space (Ho et al., 2020; Kadkhodaie and Simoncelli, 2020), and the iterative denoising process effectively learns to project noisy samples back onto this manifold. To generate samples with desired characteristics, the denoiser can be guided towards specific regions of the manifold through classifier-based or embedding-based guidance signals (Dhariwal and Nichol, 2021). However, unlike variational autoencoders or GANs (Goodfellow et al., 2014), diffusion models do not specify or infer an explicit latent representation.

By using conditional diffusion models as part of the VAE generative process, we can overcome the limitations of both models. The key idea is to learn the latent variables that will best guide the diffusion process towards the observations. Concretely, we learn a latent representation by augmenting the Markov chain $\mathbf{x}_{0:T}$ of the diffusion model with a latent variable $\mathbf{z}$ that guides the

denoising process (Fig. 1A, green box). This results in a joint distribution $p(\mathbf{x}_{0:T}, \mathbf{z})$ and approximate posterior $q(\mathbf{x}_{1:T}, \mathbf{z}|\mathbf{x}_0)$ that can be factorized as

$$p(\mathbf{x}_{0:T}, \mathbf{z}) = p_{\theta,\phi}(\mathbf{x}_{0:T}|\mathbf{z})p(\mathbf{z})$$
$$q(\mathbf{x}_{1:T}, \mathbf{z}|\mathbf{x}_0) = q_{\mathbf{x}}(\mathbf{x}_{1:T}|\mathbf{x}_0)q_\phi(\mathbf{z}|\mathbf{x}_0),$$

where the latter holds because the noisy samples in $\mathbf{x}_{1:T}$ are independent of $\mathbf{z}$ when conditioned on $\mathbf{x}_0$. Here, $q_\phi(\mathbf{z}|\mathbf{x}_0)$ is the inference network, used to infer latent $\mathbf{z}$ from observations $\mathbf{x}_0$, while $q_{\mathbf{x}}(\mathbf{x}_{1:T}|\mathbf{x}_0)$ is the (known) "posterior" that defines the conditional forward process of the diffusion model. As is common for VAEs, we assume that the model prior is an isotropic Gaussian, $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Incorporating these into the ELBO yields a joint objective function of the form:

$$\begin{aligned}
\mathcal{L}_{\text{DiVA}} &= -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_0)}\mathbb{E}_{q_{\mathbf{x}}(\mathbf{x}_{1:T}|\mathbf{x}_0,\mathbf{z})}\left[\log \frac{p_{\theta,\phi}(\mathbf{x}_{0:T}|\mathbf{z})p(\mathbf{z})}{q_{\mathbf{x}}(\mathbf{x}_{1:T}|\mathbf{x}_0)q_\phi(\mathbf{z}|\mathbf{x}_0)}\right] \\
&= -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_0)}\mathbb{E}_{q_{\mathbf{x}}(\mathbf{x}_{1:T}|\mathbf{x}_0,\mathbf{z})}\left[\log \frac{p_{\theta,\phi}(\mathbf{x}_{0:T}|\mathbf{z})}{q_{\mathbf{x}}(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_0)}\left[\log \frac{p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x}_0)}\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_0)}\left[D_{\text{KL}}\left(q_{\mathbf{x}}(\mathbf{x}_{1:T}|\mathbf{x}_0)\|p_{\theta,\phi}(\mathbf{x}_{0:T}|\mathbf{z})\right)\right] + D_{\text{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}_0)\|p(\mathbf{z})\right), \qquad (4)
\end{aligned}$$

where the second term is equivalent to the standard regularization term in the VAE objective that encourages the expected approximate posterior distribution to be close to the latent prior. Inputs $\mathbf{x}_0$ to the inference network return the corresponding mean and covariance $\boldsymbol{\mu}_\phi(\mathbf{x}_0)$ and $\Sigma_\phi(\mathbf{x}_0)$ of the posterior distribution $q_\phi(\mathbf{z}|\mathbf{x}_0)$, which are then used to analytically calculate the KL divergence term.

Implementing the generative model as conditional diffusion. The first term in DiVA's objective (Eq. 4) encourages the latent conditioned generative model $p_{\theta,\phi}(\mathbf{x}_{0:T}|\mathbf{z})$ to be close to the distribution over forward noising paths $q_{\mathbf{x}}$. Following a similar derivation to Ho et al. (2020), this term can be written as the expected mean squared error between the true and estimated noise across noise levels t:

$$\mathcal{L}_{\text{DiVA}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_0),\boldsymbol{\epsilon},t}\left[\lambda_t\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\theta,\phi}(\mathbf{x}_t, t, \mathbf{z})\|^2\right] + D_{\text{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}_0)\|p(\mathbf{z})\right), \qquad (5)$$

where $\boldsymbol{\epsilon}$ is the true noise, $\boldsymbol{\epsilon}_{\theta,\phi}$ is the function estimating it, and λ_t is a scalar hyperparameter that depends on t. For the first term in the objective, we can use Miyasawa's formula to relate the conditional MMSE estimate to the guidance score (full derivation in Appendix A.1). Using the same variance-preserving noise distribution as in Eq. (3), we see that

$$\hat{\boldsymbol{\epsilon}}_{\theta,\phi}(\mathbf{x}_t, \mathbf{z}) = \underset{\boldsymbol{\epsilon}_{\theta,\phi}}{\operatorname{argmin}} \|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_{\theta,\phi}(\mathbf{x}_t, t, \mathbf{z})\|^2 = -\sqrt{1-\bar{\alpha}_t}\,\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{z}). \qquad (6)$$

Critically, Bayes' rule can be used to express the guidance score as the sum of two terms, an unconditional denoiser score and a guidance score:

$$\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{z}) = \nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t) + \nabla_{\mathbf{x}_t}\log p(\mathbf{z}|\mathbf{x}_t). \qquad (7)$$

Plugging this back into Eq. (6), we see that the MMSE solution can be expressed as

$$\begin{aligned}
\hat{\boldsymbol{\epsilon}}_{\theta,\phi}(\mathbf{x}_t, \mathbf{z}) &= -\sqrt{1-\bar{\alpha}_t}\big(\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t) + \nabla_{\mathbf{x}_t}\log p(\mathbf{z}|\mathbf{x}_t)\big) \\
&= \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t) - \gamma_t\,\mathbf{g}_t(\mathbf{x}_t, \mathbf{z}), \qquad (8)
\end{aligned}$$

where $\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t) = -\sqrt{1-\bar{\alpha}_t}\,\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t)$ is the MMSE solution associated with an unconditional diffusion model (Eq. 3), $\mathbf{g}_t(\mathbf{x}_t, \mathbf{z}) = \nabla_{\mathbf{x}_t}\log p(\mathbf{z}|\mathbf{x}_t)$ is the guidance score, and $\gamma_t = \sqrt{1-\bar{\alpha}_t}$ is a noise level-dependent scalar term that weights the two terms appropriately.

Computing the guidance score. From the perspective of the unconditional diffusion model, the VAE posterior has a dual interpretation as a likelihood function $\mathscr{L}_\phi(\mathbf{x}_t; \mathbf{z})$, which is a function of the noisy image $\mathbf{x}_t$ over the latent domain $\mathcal{Z}$. The guidance score is the score of the likelihood evaluated at a particular latent value $Z = \mathbf{z}$, which in our case we take to be the latent corresponding to the clean image, since our goal is to reconstruct the original observation. To compute this, we take samples of the clean image latent $\mathbf{z}(\mathbf{x}_0) \sim q_\phi(\mathbf{z}|\mathbf{x}_0)$ in accordance with the expectation in Eq. (5) and compute its log likelihood, $\log \mathscr{L}_\phi(X_t = \mathbf{x}_t; Z = \mathbf{z}(\mathbf{x}_0))$. Taking the derivative of this scalar quantity with respect to the noisy image $\mathbf{x}_t$ (via auto-diff) then gives the approximate guidance score

$$\mathbf{g}_{t,\phi}(\mathbf{x}_t, \mathbf{z}) = \mathbb{E}_{\mathbf{z}'\sim q_\phi(\mathbf{z}|\mathbf{x}_0)}\big[\nabla_{\mathbf{x}_t}\log q_\phi(\mathbf{z}'|\mathbf{x}_t)\big] \qquad (9)$$

The reuse of the inference network to compute the log likelihood removes the need for a separate neural network to map the representation $\mathbf{z}$ into a guiding signal, and ties the representations of clean and noisy versions of an image together. Plugging this back into Eq.(8), and this in turn into Eq. (5), we see that the DiVA objective can be written as the influence of two separate networks in our model, the unconditional denoiser parameterized by θ, and the inference network parameterized by ϕ:

$$\mathcal{L}_{\beta\text{-DiVA}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_0),\boldsymbol{\epsilon}_t,t}\Big[\lambda_t \|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) + \gamma_t\, \mathbf{g}_{t,\phi}(\mathbf{x}_t, \mathbf{z})\|^2\Big] + \beta\, D_{\mathrm{KL}}\big(q_\phi(\mathbf{z}|\mathbf{x}_0)\|p(\mathbf{z})\big), \quad (10)$$

where β is a hyperparameter that balances the effective contribution of the KL regularization on the latent space. This is akin to the Lagrange multiplier that controls the reconstruction and regularization terms in β-VAEs (Higgins et al., 2017). Note that if we rewrite this derivation in terms of $\mathbf{x}_0$ estimates instead of $\boldsymbol{\epsilon}$ estimates, we can show, using the relation $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}$, that $\hat{\mathbf{x}}_0(\mathbf{x}_t, \mathbf{z}) = \hat{\mathbf{x}}_0(\mathbf{x}_t) + {}^{1-\bar{\alpha}_t}\!/_{\sqrt{\bar{\alpha}_t}}\mathbf{g}_t(\mathbf{x}_t, \mathbf{z})$, which reflects the additive relation between the denoising score and guidance score in Fig. 1C, $\mathcal{X}$ space. The expression that details the effect of the guidance score on the DDPM reverse transition operator is given in Appendix A.2.

Inference and guidance. Once DiVA is trained using Eq. (10), inference follows the general VAE prescription where latent samples are drawn from the posterior distribution $q_\phi(\mathbf{z}|\mathbf{x}_0)$ via the reparameterization trick. Generating image samples that resemble the given "guidance" datapoint $\mathbf{x}_0$ involves conditioning the diffusion process on draws from the posterior distribution. As in DDPMs, we start the conditional generation process with a sample from an isotropic gaussian $\mathbf{x}_t \sim \mathcal{N}(0, I)$. At each noise level t, we estimate the noise present in the noisy image via the denoiser $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$, and compute the guidance score $\mathbf{g}_{t,\phi}$ using Eq. (9). Note that while the clean image posterior remains the same at every noise level, the guidance score changes because we evaluate the clean image latent under the noisy image posterior, which changes with the noise level. The resulting guidance score biases the reverse process according to Eq. (8), such that as the noise level goes to 0, we arrive at a sample from the conditional distribution $\hat{\mathbf{x}}_0 \sim p_{\theta,\phi}(\mathbf{x}_0|\mathbf{z})$. Algorithms are provided in Appendix A.3.

Geometric intuition. In the image space, conditional generation update steps (purple arrow in Fig. 1C, $\mathcal{X}$ space) can be thought of as a combination of two forces: starting from a noisy image $\mathbf{x}_t$, attractor dynamics push the state towards the image manifold (blue arrow) while the latent-based guidance selects specific regions of the manifold (red arrow). The addition of these two forces direct the network towards regions of the data manifold that are close to the target $\mathbf{x}_0$ (green shade).

In the latent space (Fig. 1C, $\mathcal{Z}$), similarity between the current state $\mathbf{x}_t$ and the conditioning image $\mathbf{x}_0$ (as defined by the likelihood) is the Mahalanobis distance between the mean estimate of the network state in latent space $\boldsymbol{\mu}_\phi(\mathbf{x}_t)$ and the latent corresponding to the conditioning image $\mathbf{z}'(\mathbf{x}_0)$, computed under a metric that is given by the inverse of the covariance Σ_ϕ^{-1} (concentric ellipsoids). Conditional generation can thus be cast as a constrained optimization process, where the network minimizes the latent Mahalanobis distance subject to the constraints posed by the geometry of the latent space and the image manifold (for further details, see Appendix A.4).

During the learning process, the network learns to map noisy version of the same image close together in $\mathcal{Z}$ space, while accounting for uncertainty introduced by the noise in the associated posterior covariance, $\Sigma_\phi(\mathbf{x}_t)$. This effectively means learning which directions in latent space correspond to noise-induced variations versus meaningful semantic content.

3 RESULTS

Synthetic disks dataset. We first investigated whether DiVA is able to learn a semantically meaningful latent space by training the network on a synthetic dataset where the ground truth latent factors and generative structure are known, and the number of latent dimensions matches that of the latent factors. The dataset is a simplified version of that used by Kadkhodaie et al. (2024), comprised of images of circles ("disks") of a fixed radius and intensity on a blank background (Fig. 2A, B). The images are fully determined by three independent generative factors: the coordinates of the disk center, c_x, c_y, and the intensity of the background, I_{bg}. All three generative factors are uniformly distributed, and the foreground intensity of the disks is held to a constant value of 0.5.

As a first measure of whether DiVA learns a useful representation, we tested the model's ability to faithfully reconstruct images when conditioned on a given guidance image (Fig. 2C; see Appendix A.5

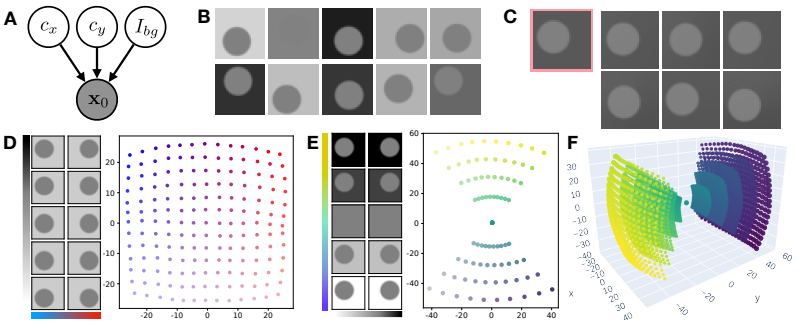

Figure 2: **Disks dataset. A)** Graphical model for the disks dataset, with coordinates of disk center (c_x, c_y) and background intensity I_b as latents. **B)** Random draws from the ground truth generative process. **C)** Samples drawn from the model, conditioned on leftmost image. **D)** Posterior means for a grid of test c_x and c_y ground truth positions, fixed I_b. **E)** Same as D, but c_y fixed during interpolation of the other two factors. **F)** Same as D, for all three factors.

for architecture details). The conditioned image generation is sensible: it preserves background intensity (which can be accurately estimated from the conditioning image), with small variations in the location of the disc reflecting some posterior uncertainty in that dimension. To quantify the latent variable's contribution in "explaining away" the observation, we measured the reduction in MSE between generated and guidance images compared to the expected MSE of the dataset with respect to the mean image, where MSE serves as a proxy for the entropy in the distribution of generated samples. Since the dataset is symmetric around the mean image, the expected MSE with respect to this image gives us a heuristic measure of the total variance of the entire dataset. We found a substantial decrease in the MSE from 5.6×10^{-1} to 4×10^{-3} when the generation is conditioned on the latent variable, indicating that the representation captures most of the variability in the dataset.

To directly probe the semantic structure of the learned representation, we systematically varied two of the generative factors and extracted the corresponding posterior means provided by the inference network. When varying the x, y coordinates of the disk, we found an approximately Cartesian representation in the latent space (Fig. 2D), while varying the background intensity and the x coordinate resulted in an approximately polar representation (Fig. 2E). The singular point in the center corresponds to an image where the background intensity exactly matches the foreground intensity, so the x, y coordinates are unspecified. The shrinking of the spatial encoding with contrast is a reflection of increasing uncertainty in position pulling the posterior mean towards the prior. We see this more clearly in the 3d space (Fig. 2F), in which we linearly interpolated along all three generative factors. Importantly, the orthogonality of the semantic axes directly reflects the independence of the three generative factors. The constraints imposed by the KL regularization term suffice for the model to learned a factorized, semantically meaningful representation of the data.

CelebA dataset. Next, we applied our method to CelebA, a dataset of diverse, high-quality images of human faces with structured attributes (e.g. facial features, expressions, and accessories) that are qualitatively discernible by human observers, which provides sematic structure that we can post-hoc evaluate in the learned representations. Following Ho et al. (2020), we used a standard UNet architecture for the denoiser, but without self-attention layers, so as to focus on the algorithmic effects. The noise level was encoded using a sinusoidal position embedding, provided as an additional input channel (Nichol and Dhariwal, 2021). For the inference network, we used a half-UNet architecture that retains the Down and Mid, but not the Up blocks. The mean and covariance of the posterior are linearly decoded from this network. Unlike Mittal et al. (2023), our inference network does not receive noise level information, which we found to be unnecessary for learning a good representation. As is common in VAEs, we restricted the posterior covariance to be diagonal, $\Sigma_\phi = \text{diag}(\boldsymbol{\sigma}_\phi^2)$. Additional architectural details are provided in Appendix A.5.

We trained the joint model from scratch using 60,000 images from the CelebA dataset, and evaluated the learned representation by assessing conditioned reconstruction quality on images from the test set. If the learned latents are useful, they should encode the pertinent features in the original guiding image that, when used to condition the generation process, reduces the MSE between itself and the generated images. Note that proper conditional generation is not trivial: since guidance scores are

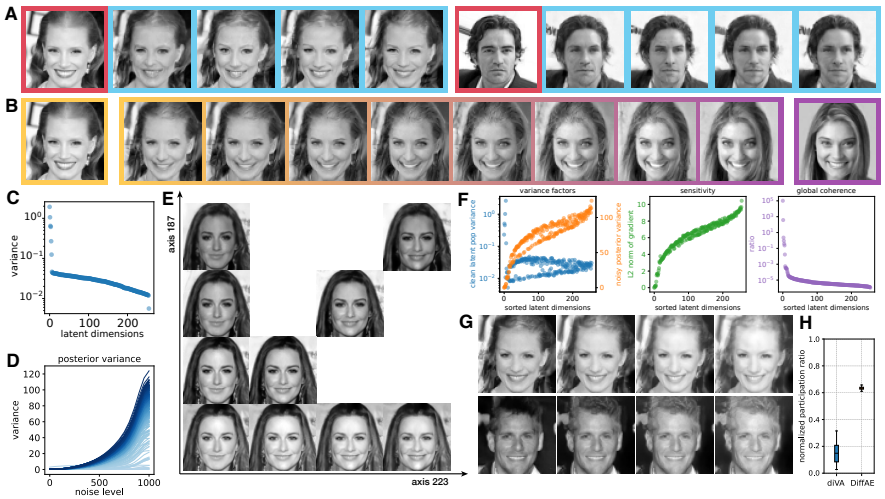

Figure 3: **CelebA. A)** Conditional image generation; conditioning image – red, samples – blue. **B)** Linear traversal between latents of leftmost (yellow) and rightmost (purple) images. **C)** The spectrum of the posterior covariance for one example image from the test set. **D)** The posterior variance as a function of noise level; color intensity marks the ordering of the axes at the largest noise level. **E)** Moving along the linear combination of two latent axes with identified semantics. **F)** Geometric characterization of latent representation. Left: across-test data variability in posterior mean (blue) and posterior variance of a noisy image (orange) for each latent dimension. Middle: norm of the sensitivity of posterior variance to input changes. Right: Global coherence of each axis. The x axes of all three plots are sorted by global coherence. **G)** Example transformation of two different images along a latent axis identified as global. **H)** Weight sparsity of binary classifiers compares how closely the semantics of latent axes align to supervised labels for DiVA vs. DiffAE.

added to the denoiser estimates in pixel space, conditioning can easily worsen generation if the latents are non-informative or misaligned with the image prior.

The generated images possess many of the same features as the guiding image, including skin tone, light position, pose, and location and shape of facial features such as the eyes, hairline, and mouth (Fig. 3A). As with the disks, we observe variability in conditionally generated image features, stemming from uncertainty in the posterior and the stochastic nature of the diffusion process. To measure sample fidelity and diversity, we computed the FID based on 10,000 images (Heusel et al., 2017). DiVA's score of 16.25 beats other diffusion-based representation learning models such as DiffAE and InfoDiffusion (Preechakul et al., 2022; Wang et al., 2023) (see Appendix A.7 for details).

To investigate the smoothness of the latent embedding we used latent traversal, linearly interpolating between the latent representations corresponding to two images from the test set (Fig. 3B). At regular intervals along this interpolant, we generated sample images conditioned on the corresponding latent. Given two endpoint images (enclosed in yellow and purple boxes in Fig. 3B), the intermediate images display semantic characteristics that change smoothly between those of the endpoint images, indicating that the learned latent space is smooth and continuous.

Factorized semantics. The results on the disk images suggest DiVA latents can capture independent generative factors in the data. Moreover, our explicitly probabilistic approach point to the structure of the posterior covariance as a natural lens into the interpretability of individual latent axes.

Empirically, the posterior variance for any given image exhibits a few dominant modes alongside many smaller ones (Fig. 3C), indicating that, within the local neighborhood of the mean latent representation, certain latent axes have significantly higher uncertainty. Adding noise to the image increases variance across these axes (Fig. 3D), suggesting that posterior variance reflects the inference network's uncertainty about the corresponding features. We hypothesize that axes with low posterior variance at low noise levels encode robust semantic features, while the rate of variance increase with noise level indicates the spatial scale of the encoded features. Specifically, large-scale features, still prominent at relatively high noise levels, should correspond to axes with low uncertainty even under significant noise, whereas small-scale features will exhibit overall higher uncertainty.

To test this, we perturbed a clean test image's latent representation along axes with low posterior variance. The resulting images showed consistent changes in specific semantic attributes, suggesting that the model factorizes representations semantically (Fig. 3E, smiling horizontally, forward lighting vertically). Interpolating along a linear combination of two axes produced images reflecting both attributes, suggesting a combinatorial code (Fig. 3E, diagonal). Meanwhile, axes that displayed large changes in posterior variance tended to correspond to finer scale features, such as movements of the eyes and mouth. Finally, we measured disentanglement using Total AUROC Difference (TAD) (Yeats et al., 2022). When trained on CelebA, DiVA attains a TAD score of 0.583, capturing 3 attributes, which outperforms InfoDiffusion (TAD of 0.299 with 3 attributes captured), and significantly outperforms other baseline generative models such as β-VAE (full comparisons in Appendix A.7). These findings reveal that DiVA's latent space is structured to disentangle semantic features along specific axes, offering a foundation for interpretable and controllable generative modeling.

How global is the mapping between latent axes and semantic meaning? While perturbation along individual axes is interpretable in the neighborhood of the conditioning image, we found that in general this mapping is not preserved across the full latent space. To identify the axes whose semantic meaning is preserved globally, we derived a formula that provides sufficient conditions for global axes under additive transformations by assessing the $\mathbf{z}$-independence of the linear local approximation of the inference network (see Appendix A.8). When applying this metric (Fig. 3F, right) we find the axes with the largest measure to be "junk" dimensions that encode the pixel-level noise, whose semantic mapping is in effect global. However, most of the identified global axes are interpretable, reflecting e.g. global lightning effects (Fig. 3G).

Alignment with explicit labels. A separate way to assess the semantic structure of the representation is by measuring the alignment between latent axes and recognizable semantic labels. To this end, we trained separate logistic regression classifiers on the learned representation to predict each of the 40 binary semantic labels. If the axes are well aligned with the supervised attributes, the weights of the classifier should be sparse, as only a small subset of the axes need to be recruited to form the decision boundary. We used the normalized participation ratio (PR) of each classifier's coefficients as a measure of sparsity; this is 1 if only one latent axis is used for a decision, and it becomes the square root of the dimensionality of the latent space if all axes are used equally. The normalized PR is generally low for DiVA, in particular when compared to state of the art alternative DiffAE (Fig 3H), suggesting that our approach is superior in factorizing the latent space into useful semantic axes.

Figure 4: **Unsupervised extraction of latent representation from a pre-trained diffusion model. A)** Top: Samples from the trained model, conditioned on leftmost image (green). Bottom: samples drawn from the unconditional diffusion model. **B)** Straightness of latent trajectories over the course of one naturalistic video from CelebV. **C)** Average straightness of latent encoding (over 50 naturalistic videos) for DiVA, compared to trajectories in pixel space and latent trajectories of DiffAE.

Feature extraction for pretrained diffusion models. Since the denoising and guidance terms are parameterized by separate neural networks, one might sensibly wonder if training them jointly is strictly necessary or if the inference network could be trained separately on top of a pre-trained denoiser. To test this, we used a DDPM from the Huggingface repository pretrained on CelebA-HQ-256 (Ho et al., 2020; Karras et al., 2018) and then trained the inference network to guide the responses of the denoiser towards the corresponding clean image using the same objective as before. We found that samples drawn from the conditioned distribution exhibit similar semantic characteristics as the guiding image, such as lighting direction, facial expression, and hairline, while remaining a naturalistic image with high frequency details (Fig. 4). As before, the conditioning reduces variability compared to samples from the unconditional model (Fig. 4A, second row), but

there is still appreciable diversity among the samples. Overall, separate training of the conditioning network seems to generally preserve the properties seen with joint training with a substantial reduction in training effort.

Encoding of video trajectories. One property that makes for good latent representations is their predictability over time, measured by the ability to represent temporal sequences in a manner that supports linear extrapolation (Niu et al., 2024). Since movement along individual latent axes corresponds to naturalistic transformations in images (Fig. 3E), we hypothesized that DiVA may map natural movements in naturalistic videos, such as a mouth opening and closing, to relatively straight trajectories in latent space. To test this, we encoded frames from a naturalistic facial attributes video dataset (Yu et al., 2023) into latent representations using the DiVA model for CelebA-HQ. Following Hénaff et al. (2019), we quantified the "straightness" of temporal sequences in both latent and pixel domains by computing discrete curvature, defined as the cosine similarity between vectors connecting consecutive pairs of frames in the sequence: $s = \mathbf{d}_t \cdot \mathbf{d}_{t+1} / (\|\mathbf{d}_t\| \|\mathbf{d}_{t+1}\|)$, where $\mathbf{d}_t = \mathbf{f}_{t+1} - \mathbf{f}_t$ represents the difference between consecutive frames, or latent posterior means. Despite being trained solely on static images, our model encodes frames of these videos such that their trajectories in the 512-dimensional latent space exhibit near-unity cosine similarity across large portions of a given video (Fig. 4B). Compared to pixel-domain trajectories and those from other diffusion-based representation learning models (Preechakul et al., 2022), DiVA produces significantly straighter latent trajectories. This effect holds consistently across 50 randomly sampled videos from the dataset (Fig. 4C). Thus, by leveraging its semantically factorized representation, DiVA naturally straightens dynamic patterns without temporal supervision.

4 Discussion

While recent work has demonstrated the promise of unsupervised representation learning with diffusion models (Fuest et al., 2024; Preechakul et al., 2022; Hudson et al., 2024; Mittal et al., 2023; Wang et al., 2023; Zhang et al., 2022; Yang et al., 2023; Kim et al., 2024; Li et al., 2024), past approaches typically rely on (often deterministic) encoders that approximate the log-likelihood, or use objective functions that are special cases of the ELBO with extra regularization terms (see Appendix A.6 for detailed comparison to prior methods). DiVA diverges from this paradigm by optimizing an exact ELBO that encodes both clean and noisy images simultaneously. This dual encoding strategy introduces a novel form of implicit regularization. By requiring the inference network to predict the log-likelihood of clean image features while conditioned on noisy observations, DiVA learns to effectively separate signal from noise and calibrate its posterior variance to capture the irreducible uncertainty inherent in the data. The probabilistic nature of these paired observations constrains the latent space geometry, promoting representations where semantically similar images cluster together. Our empirical analysis reveals that this approach successfully mitigates the "latent holes" problem that commonly afflicts variational autoencoders (Rezende and Viola, 2018; Falorsi et al., 2018; Li et al., 2021), producing perceptually smoother and more coherent latent spaces.

Pairing a powerful recognition model with strong factorization Higgins et al. (2017) assumptions on the prior encourages disentanglement without sacrificing generative performance. The KL regularization term enforces an information bottleneck by encouraging posterior whitening, which promotes maximum entropy representations in the latent space Burgess et al. (2018). Crucially, the conditional denoising objective encourages the latents to capture meaningful structure in the observed data by leveraging the implicit image prior learned by the diffusion model. This approach sidesteps the traditional rate-distortion trade-off in VAEs: rather than compromising reconstruction fidelity to achieve better disentanglement, DiVA maintains perceptually high-quality generations through the expressive power of the diffusion prior, even when the latent representation is heavily regularized. This information theoretic perspective suggests interesting connections to the I-MMSE relation (Guo et al., 2005; Kong et al., 2023), which has recently been used to link conditional diffusion scores with estimates of mutual information (Franzese et al., 2024). Such connections might provide alternative theoretical foundations for understanding representation learning in conditional diffusion models and could potentially allow for more direct optimization of information-theoretic objectives such as InfoMax (Linsker, 1988). We leave the exploration of these connections as an exciting direction for future work. Overall, our results represent an important and mathematically rigorous step towards building learning systems that extract complex semantical structure from data in an unsupervised manner while maintaining competitive generative performance.

REFERENCES

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives, April 2014.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*, February 2017.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*, 2013.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. https://arxiv.org/abs/1401.4082v3, January 2014.

Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end Optimized Image Compression. *arXiv:1611.01704 [cs, math]*, March 2017.

Johannes Ballé, Philip A. Chou, David Minnen, Saurabh Singh, Nick Johnston, Eirikur Agustsson, Sung Jin Hwang, and George Toderici. Nonlinear Transform Coding. *IEEE Journal of Selected Topics in Signal Processing*, 15(2):339–353, February 2021. ISSN 1941-0484. doi: 10.1109/JSTSP.2020.3034501. URL https://ieeexplore.ieee.org/document/9242247/.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, 2000. URL https://arxiv.org/abs/physics/0004057.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep Variational Information Bottleneck, October 2019.

Vedant Sachdeva, Thierry Mora, Aleksandra M. Walczak, and Stephanie Palmer. Optimal prediction with resource constraints using the information bottleneck. Preprint, Biophysics, May 2020.

Ferenc Csikor, Balázs Meszéna, Bence Szabó, and Gergő Orbán. Top-down inference in an early visual cortex inspired hierarchical Variational Autoencoder, June 2022.

Hadi Vafaii, Jacob L. Yates, and Daniel A. Butts. Hierarchical VAEs provide a normative account of motion processing in the primate brain. Preprint, Neuroscience, September 2023.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational Inference of Disentangled Latent Concepts from Unlabeled Observations, December 2018. URL http://arxiv.org/abs/1711.00848. arXiv:1711.00848 [cs].

Harshvardhan Sikka, Weishun Zhong, Jun Yin, and Cengiz Pehlevan. A Closer Look at Disentangling in β-VAE. *arXiv:1912.05127 [cs, stat]*, December 2019.

Laura Manduchi, Moritz Vandenhirtz, Alain Ryser, and Julia Vogt. Tree Variational Autoencoders. *Advances in Neural Information Processing Systems*, 36:54952–54986, December 2023.

Alexej Klushyn, Nutan Chen, Richard Kurle, Botond Cseke, and Patrick van der Smagt. Learning Hierarchical Priors in VAEs. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Emile Mathieu, Tom Rainforth, N. Siddharth, and Yee Whye Teh. Disentangling Disentanglement in Variational Autoencoders, June 2019.

Wei Cheng, Gregory Darnell, Sohini Ramachandran, and Lorin Crawford. Generalizing Variational Autoencoders with Hierarchical Empirical Bayes, July 2020.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1530–1538, Lille, France, July 2015. JMLR.org.

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders, April 2019.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. InfoVAE: Information Maximizing Variational Autoencoders, May 2018.

Antoine Wehenkel and Gilles Louppe. Diffusion Priors In Variational Autoencoders, June 2021.

Michal Rolínek, Dominik Zietlow, and Georg Martius. Variational Autoencoders Pursue PCA Directions (by Accident). In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12398–12407, Long Beach, CA, USA, June 2019. IEEE. ISBN 978-1-7281-3293-8. doi: 10.1109/CVPR.2019.01269.

Chris Cremer, Xuechen Li, and David Duvenaud. Inference Suboptimality in Variational Autoencoders, May 2018. URL http://arxiv.org/abs/1801.03558. arXiv:1801.03558 [cs].

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*, October 2020.

Michael Fuest, Pingchuan Ma, Ming Gui, Johannes Schusterbauer, Vincent Tao Hu, and Bjorn Ommer. Diffusion Models and Representation Learning: A Survey, June 2024. URL http://arxiv.org/abs/2407.00783. arXiv:2407.00783 [cs].

Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021.

Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, December 2021.

Muthu Chidambaram, Khashayar Gatmiry, Sitan Chen, Holden Lee, and Jianfeng Lu. What does guidance do? A fine-grained analysis in a simple setting, September 2024.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: Unleashing the Diversity of Diffusion Models through Condition-Annealed Sampling, May 2024. URL http://arxiv.org/abs/2310.17347. arXiv:2310.17347 [cs].

Tim Kaiser, Nikolas Adaloglou, and Markus Kollmann. The Unreasonable Effectiveness of Guidance for Diffusion Models, November 2024. URL http://arxiv.org/abs/2411.10257. arXiv:2411.10257 [cs] version: 1.

Tariq Berrada Ifriqi, Pietro Astolfi, Melissa Hall, Reyhane Askari-Hemmat, Yohann Benchetrit, Marton Havasi, Matthew Muckley, Karteek Alahari, Adriana Romero-Soriano, Jakob Verbeek, and Michal Drozdzal. On Improved Conditioning Mechanisms and Pre-training Strategies for Diffusion Models, January 2025. URL http://arxiv.org/abs/2411.03177. arXiv:2411.03177 [cs].

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance Weighted Autoencoders, November 2016. URL http://arxiv.org/abs/1509.00519. arXiv:1509.00519 [cs].

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Sø ren Kaae Sø nderby, and Ole Winther. Ladder Variational Autoencoders. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Danilo Jimenez Rezende and Fabio Viola. Taming VAEs, October 2018.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265. PMLR, June 2015.

Zahra Kadkhodaie and Eero P. Simoncelli. Solving Linear Inverse Problems Using the Prior Implicit in a Denoiser, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

Diederik P. Kingma and Ruiqi Gao. Understanding Diffusion Objectives as the ELBO with Simple Data Augmentation, September 2023.

Herbert Robbins. An Empirical Bayes Approach to Statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 3.1, pages 157–164. University of California Press, January 1956.

Koichi Miyasawa et al. An empirical bayes estimator of the mean of a normal population. *Bull. Inst. Internat. Statist*, 38(181-188):1–2, 1961.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks, June 2014. URL http://arxiv.org/abs/1406.2661. arXiv:1406.2661 [stat].

Z Kadkhodaie, F Guth, E P Simoncelli, and S Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representation. In *Int'l Conf on Learning Representations (ICLR)*, volume 12, Vienna, Austria, May 2024.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8162–8171. PMLR, July 2021.

Sarthak Mittal, Korbinian Abstreiter, Stefan Bauer, Bernhard Schölkopf, and Arash Mehrjou. Diffusion Based Representation Learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 24963–24982. PMLR, July 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10619–10629, 2022.

Yingheng Wang, Yair Schiff, Aaron Gokaslan, Weishen Pan, Fei Wang, Christopher De Sa, and Volodymyr Kuleshov. InfoDiffusion: Representation Learning Using Information Maximizing Diffusion Models, June 2023.

Eric Yeats, Frank Liu, David Womble, and Hai Li. NashAE: Disentangling Representations Through Adversarial Covariance Minimization. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, volume 13687, pages 36–51. Springer Nature Switzerland, Cham, 2022. ISBN 978-3-031-19811-3 978-3-031-19812-0. doi: 10.1007/978-3-031-19812-0_3.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation, February 2018. URL http://arxiv.org/abs/1710.10196. arXiv:1710.10196 [cs].

Xueyan Niu, Cristina Savin, and Eero P. Simoncelli. Learning predictable and robust neural representations by straightening image sequences. *Advances in Neural Information Processing Systems*, 37:40316–40335, December 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/473c578564ed9fe8041abfef772d37db-Abstract-Conference.html.

Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. CelebV-Text: A Large-Scale Facial Text-Video Dataset, March 2023. URL http://arxiv.org/abs/2303.14717. arXiv:2303.14717 [cs].

Olivier J. Hénaff, Robbe L. T. Goris, and Eero P. Simoncelli. Perceptual straightening of natural videos. *Nature Neuroscience*, 22(6):984–991, June 2019.

Drew A. Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K. Lampinen, Andrew Jaegle, James L. McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. SODA: Bottleneck Diffusion Models for Representation Learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23115–23127, Seattle, WA, USA, June 2024. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.02181.

Zijian Zhang, Zhou Zhao, and Zhijie Lin. Unsupervised Representation Learning from Pre-trained Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 35:22117–22130, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/8aff4ffcf2a9d41692a805b3987e29ea-Abstract-Conference.html.

Tao Yang, Yuwang Wang, Yan Lu, and Nanning Zheng. DisDiff: Unsupervised Disentanglement of Diffusion Probabilistic Models. In *Thirty-Seventh Conference on Neural Information Processing Systems*, November 2023.

Yeongmin Kim, Kwanghyeon Lee, Minsang Park, Byeonghu Na, and Il-chul Moon. Diffusion Bridge AutoEncoders for Unsupervised Representation Learning. In *The Thirteenth International Conference on Learning Representations*, October 2024.

Tianhong Li, Dina Katabi, and Kaiming He. Return of Unconditional Generation: A Self-supervised Representation Generation Method. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, November 2024.

Luca Falorsi, Pim de Haan, Tim R. Davidson, Nicola De Cao, Maurice Weiler, Patrick Forré, and Taco S. Cohen. Explorations in Homeomorphic Variational Auto-Encoding, July 2018. URL http://arxiv.org/abs/1807.04689. arXiv:1807.04689 [stat].

Ruizhe Li, Xutan Peng, and Chenghua Lin. On the Latent Holes of VAEs for Text Generation, October 2021. URL http://arxiv.org/abs/2110.03318. arXiv:2110.03318 [cs].

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv:1804.03599 [cs, stat]*, April 2018.

Dongning Guo, S. Shamai, and S. Verdu. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, April 2005. ISSN 1557-9654. doi: 10.1109/TIT.2005.844072.

Xianghao Kong, Rob Brekelmans, and Greg Ver Steeg. Information-Theoretic Diffusion, February 2023.

Giulio Franzese, Mustapha Bounoua, and Pietro Michiardi. MINDE: Mutual Information Neural Diffusion Estimation, May 2024.

R. Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988. doi: 10.1109/2.36.

H. Helmholtz. LXXXI. On the theory of compound colours. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 4(28):519–534, January 1852. ISSN 1941-5982. doi: 10.1080/14786445208647175. URL https://doi.org/10.1080/14786445208647175. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14786445208647175.

Song Chun Zhu, Yingnian Wu, and David Mumford. Filters, Random Fields and Maximum Entropy (FRAME): Towards a Unified Theory for Texture Modeling. *International Journal of Computer Vision*, 27(2):107–126, March 1998. ISSN 1573-1405. doi: 10.1023/A:1007925832420. URL https://doi.org/10.1023/A:1007925832420.

Javier Portilla and Eero P. Simoncelli. A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *International Journal of Computer Vision*, 40(1):49–70, October 2000. ISSN 1573-1405. doi: 10.1023/A:1026553619983. URL https://doi.org/10.1023/A:1026553619983.

Jeremy Freeman and Eero P. Simoncelli. Metamers of the ventral stream. *Nature Neuroscience*, 14(9):1195–1201, September 2011. ISSN 1546-1726. doi: 10.1038/nn.2889. URL https://www.nature.com/articles/nn.2889. Publisher: Nature Publishing Group.

Aravindh Mahendran and Andrea Vedaldi. Understanding Deep Image Representations by Inverting Them, November 2014. URL http://arxiv.org/abs/1412.0035. arXiv:1412.0035 [cs].

Jenelle Feather, Guillaume Leclerc, Aleksander Mądry, and Josh H. McDermott. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, 26(11):2017–2034, November 2023. ISSN 1546-1726. doi: 10.1038/s41593-023-01442-0. URL https://www.nature.com/articles/s41593-023-01442-0. Publisher: Nature Publishing Group.

Benjamin B. Machta, Ricky Chachra, Mark K. Transtrum, and James P. Sethna. Parameter Space Compression Underlies Emergent Theories and Predictive Models. *Science*, 342(6158):604–607, November 2013.

Mark K. Transtrum, Benjamin B. Machta, Kevin S. Brown, Bryan C. Daniels, Christopher R. Myers, and James P. Sethna. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *The Journal of Chemical Physics*, 143(1):010901, July 2015.

Sreyas Mohan, Zahra Kadkhodaie, Eero P. Simoncelli, and Carlos Fernandez-Granda. Robust And Interpretable Blind Image Denoising Via Bias-Free Convolutional Neural Networks. In *International Conference on Learning Representations*, March 2022.

A APPENDIX

The anonymized code repository can be found at: https://anonymous.4open.science/r/diva-D7DA

A.1 RELATION OF MMSE SOLUTION TO CONDITIONAL SCORE

Here, we provide a derivation of how the guidance score relates to the MMSE solution for the noise present in the noisy image (Eq. 6). For a variance-preserving noising operator $p(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)I)$, Miyasawa/Tweedie's formula is given as

$$\mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_t,\mathbf{z})}[\mathbf{x}_0] = \frac{1}{\sqrt{\bar{\alpha}_t}}[\mathbf{x}_t + (1-\bar{\alpha}_t)\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{z})].$$

We can compute the expected value of the noise in the noisy image

$$\begin{aligned}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_t,\mathbf{z})}[\boldsymbol{\epsilon}] &= \int \boldsymbol{\epsilon}\, p(\mathbf{x}_0|\mathbf{x}_t,\mathbf{z})d\mathbf{x}_0 \\ &= \frac{1}{\sqrt{1-\bar{\alpha}_t}}\left[\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_t,\mathbf{z})}[\mathbf{x}_0]\right].\end{aligned}$$

where on the second line we have substituted in $\boldsymbol{\epsilon} = {}^1/_{\sqrt{1-\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0)$ via the re-parameterization trick. Now, rearranging for and substituting the expected value of $\mathbf{x}_0$ into Tweedie's formula above, we are left with

$$\begin{aligned}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_t,\mathbf{z})}[\mathbf{x}_0] &= \frac{1}{\sqrt{1-\bar{\alpha}_t}}\left[\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\left(\frac{1}{\sqrt{\bar{\alpha}_t}}[\mathbf{x}_t + (1-\bar{\alpha}_t)\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{z})]\right)\right] \\ &= \frac{1}{\sqrt{1-\bar{\alpha}_t}}\left[\mathbf{x}_t - (\mathbf{x}_t + (1-\bar{\alpha}_t)\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{z}))\right] \\ &= \frac{1}{\sqrt{1-\bar{\alpha}_t}}\left(-(1-\bar{\alpha}_t)\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{z})\right) \\ &= -\sqrt{1-\bar{\alpha}_t}\,\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{z})\end{aligned}$$

which is the relation given by Eq. 6 in the main text.

A.2 GUIDING TRANSITION OPERATORS

Here we derive the equation for how the guidance score affects the DDPM reverse transition operator, given by $p(\mathbf{x}_t|\mathbf{x}_{t+1})$. The classifier guidance equation in Eq. 7 can be expressed in terms of the transition operators:

$$\begin{aligned}\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{x}_{t+1},\mathbf{z}) &= \nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{x}_{t+1}) + \nabla_{\mathbf{x}_t}\log p(\mathbf{z}|\mathbf{x}_t,\mathbf{x}_{t+1}) \\ &= \nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{x}_{t+1}) + \nabla_{\mathbf{x}_t}\log p(\mathbf{z}|\mathbf{x}_t),\end{aligned}$$

where we have used the fact that since $\mathbf{x}_{t+1}$ is a noisier image than $\mathbf{x}_t$, it carries less information about $\mathbf{z}$ than $\mathbf{x}_t$, so the dependence on $\mathbf{x}_{t+1}$ can be dropped. In practice, this problem is equivalent to asking how to sample from $p(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{z})$ over all noise levels t. For a given noise schedule determined by $\{\alpha_t\}$, we evaluate $\mathbf{x}_t \sim p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\boldsymbol{\mu}_\theta, \sqrt{1-\alpha_t}I)$ by estimating the transition operator mean

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)\right),$$

using which we can estimate the next state in the Markov chain $\mathbf{x}_{t-1} = \boldsymbol{\mu}_\theta(\mathbf{x}_t) + \sqrt{1-\alpha_t}\,\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$. To sample from the $\mathbf{z}$-conditional distribution instead, this function must be dependent

on $\boldsymbol{\epsilon}_{\theta,\phi}(\mathbf{x}_t, \mathbf{z})$ rather than on $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)$:

$$\begin{aligned}
\boldsymbol{\mu}(\mathbf{x}_t, \mathbf{z}) &= \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\theta,\phi}(\mathbf{x}_t, \mathbf{z})\right) \\
&= \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\left(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t) - \sqrt{1-\bar{\alpha}_t}\nabla_{\mathbf{x}_t}\log q_\phi(\mathbf{z}|\mathbf{x}_t)\right)\right) \\
&= \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)\right) - \frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla_{\mathbf{x}_t}\log q_\phi(\mathbf{z}|\mathbf{x}_t) \\
&= \boldsymbol{\mu}_\theta(\mathbf{x}_t) - \frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla_{\mathbf{x}_t}\log q_\phi(\mathbf{z}|\mathbf{x}_t).
\end{aligned}$$

Since $\mathbf{x}_{t-1}(\mathbf{x}_t, \mathbf{z}) = \boldsymbol{\mu}(\mathbf{x}_t, \mathbf{z}) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}$, we can rewrite this relation in terms of the noisy states:

$$\begin{aligned}
\mathbf{x}^*_{t-1}(\mathbf{x}_t, \mathbf{z}) &= \mathbf{x}^*_{t-1}(\mathbf{x}_t) - \frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla_{\mathbf{x}_t}\log q_\phi(\mathbf{z}|\mathbf{x}_t) \\
&= \mathbf{x}^*_{t-1}(\mathbf{x}_t) - \frac{1-\alpha_t}{\sqrt{\alpha_t}}\mathbf{g}_{t,\phi}.
\end{aligned}$$

A.3 Training and conditional generation algorithms

Below, we provide algorithms for training DiVA and for performing conditional generation using the trained networks.

Algorithm 1: Conditional generation, given a clean guidance image $\mathbf{x}_0$, trained denoiser $\boldsymbol{\epsilon}_\theta$ and inference network f_ϕ

$(\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0) \leftarrow f_\phi(\mathbf{x}_0)$ `// encode guidance image`
$\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$
$\mathbf{z} \leftarrow \boldsymbol{\mu}_0 + \boldsymbol{\sigma}_0 \cdot \boldsymbol{\epsilon}$ `// sample clean image latent`

Initialize $t \leftarrow T$; $\mathbf{x}_t \sim \mathcal{N}(0, I)$
while $t \neq 0$ **do**
  $(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t) \leftarrow f_\phi(\mathbf{x}_t)$ `// encode noisy image`
  $\log q \leftarrow -1/2\big(d_M(\mathbf{z}, \boldsymbol{\mu}_t) + \log|\boldsymbol{\sigma}_t I|\big)$ `// log posterior`
  $\mathbf{g}_t \leftarrow \nabla_{\mathbf{x}_t}\log q$ `// guidance score`
  $\boldsymbol{\mu} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta\right)$
  $\Sigma \leftarrow \sqrt{1-\alpha_t}I$
  $\mathbf{x}_t \sim \mathcal{N}(\boldsymbol{\mu} + \Sigma\mathbf{g}_t, \Sigma)$ `// guidance`
  $t \leftarrow t - 1$
end

A.4 Conditional generation as constrained optimization

As mentioned in the "Geometric Intuition" subsection, we can think of conditional generation as a constrained optimization problem, in which we have multiple constraints imposed by the inference network and the denoiser. To make this perspective more concrete, consider the generation procedure, given by Algorithm 1. We start with a sample of white noise, $\mathbf{x}_t \sim \mathcal{N}(0, I)$, from which we take gradient steps towards a region of large guided score:

$$\begin{aligned}
\mathbf{x}_t &\leftarrow \mathbf{x}_t + \nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{z}) \\
&= \mathbf{x}_t + \nabla_{\mathbf{x}_t}\log q_\phi(\mathbf{z}|\mathbf{x}_t) + \nabla_{\mathbf{x}_t}\log p_\theta(\mathbf{x}_t),
\end{aligned}$$

where we have decomposed the conditional score using Bayes' rule. Since the posterior q_ϕ is assumed to be conditional Gaussian, the analytic form of the log likelihood is given as

$$\log \mathscr{L}_\phi(\mathbf{x}_t; \mathbf{z}) = -\frac{1}{2}\,\mathbb{E}_{\mathbf{z}'\sim q_\phi(\mathbf{z}|\mathbf{x}_0)}\Big[\, d_M\big(\mathbf{z}'_\phi(\mathbf{x}_0), \boldsymbol{\mu}_\phi(\mathbf{x}_t)\big) + \log|\Sigma_\phi(\mathbf{x}_t)| + c\,\Big], \tag{11}$$

Algorithm 2: Training, given a denoiser network $\boldsymbol{\epsilon}_\theta$ and inference network f_ϕ with randomly initialized weights

while *not converged* **do**

$\mathbf{x}_0 \sim p(\mathbf{x}_0)$

$t \sim [0, T]$

$\mathbf{x}_t \sim p_{\text{fwd}}(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)I)$ `// noise the image`

$\boldsymbol{\epsilon}_\mathbf{x} \leftarrow {}^{(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0)}/_{\sqrt{1-\bar{\alpha}_t}}$

$\boldsymbol{\epsilon}_\mathbf{z} \sim \mathcal{N}(0, I)$

$(\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0) \leftarrow f_\phi(\mathbf{x}_0)$ `// encode clean guidance image`

$\mathbf{z} \leftarrow \boldsymbol{\mu}_0 + \boldsymbol{\sigma}_0 \cdot \boldsymbol{\epsilon}_\mathbf{z}$

$(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t) \leftarrow f_\phi(\mathbf{x}_t)$ `// encode noisy image`

$\log q \leftarrow -1/2\big(d_M(\mathbf{z}, \boldsymbol{\mu}_t) + \log|\boldsymbol{\sigma}_t I|\big)$ `// log posterior`

$\mathbf{g}_{t,\phi}(\mathbf{z}) \leftarrow \nabla_{\mathbf{x}_t} \log q$ `// guidance score`

$\mathcal{L}_\mathbf{x} \leftarrow \|\boldsymbol{\epsilon}_\mathbf{x} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t) - \sqrt{1-\bar{\alpha}_t}\,\mathbf{g}_{t,\phi}(\mathbf{x}_t, \mathbf{z})\|^2$

$\mathcal{L}_\mathbf{z} \leftarrow D_{KL}\left(\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0)\|\mathcal{N}(0, I)\right)$

take gradient steps on $\nabla_{\theta,\phi}\left(\mathcal{L}_\mathbf{x} + \beta\mathcal{L}_\mathbf{z}\right)$

end

where $d_M = \big(\mathbf{z}'_\phi(\mathbf{x}_0) - \boldsymbol{\mu}_\phi(\mathbf{x}_t)\big)^\top \Sigma_\phi^{-1}\big(\mathbf{z}'_\phi(\mathbf{x}_0) - \boldsymbol{\mu}_\phi(\mathbf{x}_t)\big)$ is the Mahalanobis distance between the mean estimate of the network state in latent space $\boldsymbol{\mu}_\phi(\mathbf{x}_t)$ and the latent corresponding to the target image $\mathbf{z}'(\mathbf{x}_0)$, computed under a metric that is given by the inverse of the covariance Σ_ϕ^{-1} (red arrow and concentric ellipsoids in Fig. 1C, $\mathcal{X}$ space).

Using this expression for the log likelihood, it becomes clear that we are taking gradient steps with respect to multiple constraints:

$$
\begin{aligned}
\mathbf{x}_t &\leftarrow \mathbf{x}_t + \nabla_{\mathbf{x}_t} \log q_\phi(\mathbf{z}|\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t) \\
&= \mathbf{x}_t - \frac{1}{2}\nabla_{\mathbf{x}_t}\left(\mathbb{E}_{\mathbf{z}'\sim q_\phi(\mathbf{z}|\mathbf{x}_0)}\Big[d_M\left(\mathbf{z}', \boldsymbol{\mu}_\phi(\mathbf{x}_t)\right) + \log|\Sigma_\phi(\mathbf{x}_t)|\Big]\right) + \nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t) \\
&= \mathbf{x}_t - \frac{1}{2}\mathbb{E}_{\mathbf{z}'\sim q_\phi(\mathbf{z}|\mathbf{x}_0)}\Big[\nabla_{\mathbf{x}_t} d_M(\mathbf{z}', \boldsymbol{\mu}_\phi(\mathbf{x}_t))\Big] - \frac{1}{2}\nabla_{\mathbf{x}_t} \log|\Sigma_\phi(\mathbf{x}_t)| + \nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t).
\end{aligned}
$$

There are three terms, each corresponding to a specific constraint during this optimization procedure. The first constraint is to land in a region of the image space such that its representation minimizes its Mahalanobis distance to the latent of the guiding image $\mathbf{z}'(\mathbf{x}_0)$. This is related to the metamer perspective (Helmholtz, 1852; Zhu et al., 1998; Portilla and Simoncelli, 2000; Freeman and Simoncelli, 2011; Mahendran and Vedaldi, 2014; Feather et al., 2023), where the goal is to produce data samples that share the same latent representation as a target datum. Metamer generation can be formulated as a constrained optimization problem, minimizing the L2 distance between the representations of the current sample and the target, $\mathbf{x}_t \leftarrow \mathbf{x}_t - \nabla_{\mathbf{x}_t}\|\mathbf{r}_{\text{target}} - \mathbf{r}_{\mathbf{x}_t}\|^2$, where $\mathbf{r}_{\mathbf{x}_t}$ is the representation of the current sample $\mathbf{x}_t$, and $\mathbf{r}_{\text{target}}$ that the target datum. Our approach generalizes this framework by replacing the L2 distance with the log likelihood from Eq. 11. In this view, the inference network constrains the generation process to produce samples that are metameric to the guidance image as defined by the latent representation.

The next term corresponds to taking steps towards regions of the image space that minimizes the log determinant of the covariance matrix. Geometrically, we can think of this as reducing the volume of the uncertainty ellipsoid over the features identified by the inference network. In a well-calibrated Bayesian network, the posterior uncertainty over the identified features should match the irreducible uncertainty present in the image. For DiVA, the irreducible uncertainty in the latent space comes from the noise in the noisy observations, and as such we should expect that the log determinant of the covariance to correlate strongly with the variance of the noisy image.

The final term corresponds to the denoiser, parameterized by θ, which ensures that the network is driven towards the natural image manifold, i.e. the set of highly probable images.

The region of the $\mathcal{X}$ space that we land on is at the intersection of these three signals. This perspective suggests that we can increase the diversity (i.e. the entropy) of the generated images by relaxing

one or more of these constraints. For example, if the generation procedure were given only by the third term, the network would be driven to satisfy only the image prior, recovering the unconditional generation of DDPM. Alternatively, we might choose to calculate the Mahalanobis distance over only a subset of the latent dimensions, in effect making sure that $\boldsymbol{\mu}_\phi$ is close to $\mathbf{z}'$ only in those latent dimensions. This has the effect of making those latent axes "rigid", while the rest of the axes are "sloppy", since large distances in these dimensions are discounted Machta et al. (2013); Transtrum et al. (2015). Back in the pixel space, this equates to finding images that possess a subset of the semantic attributes in the original guiding image, i.e. those that correspond to the constrained feature dimensions.

A.5 Architecture and training details

Parameter	Disks	CelebA 64	CelebA-HQ 256
Inference net arch.	ConvNet	Half-UNet	Half-UNet
Infnet base channels	48	64	64
Infnet channel multiplier	[2, 2]	[1, 1, 1, 1, 1, 1]	[2, 2, 2, 2, 1, 1]
Infnet nonlinearity	ReLU	ReLU	SiLU
Denoiser arch.	UNet (no attn)	UNet (no attn)	UNet (with attn, frozen)
Denoiser base channels	128	128	128
Denoiser ch. multiplier	[1, 1, 1, 1, 1, 1]	[1, 1, 1, 1, 1, 1]	[1, 1, 2, 2, 4, 4]
Num. noise levels	400	1000	1000
Latent dimensionality	3	256	512
KL weight	5e-6	1e-3	1e-8
Noise schedule	Linear	Cosine	Linear
Training set size	2000	60,000	30,000
Batch size	512	512	512
Num. epochs trained	1000	5000	70
Learning rate	6e-3	1e-3	3e-5
Timestep sampling dist.	Uniform	Uniform	Monotonically increasing
Resources (num. A100)	1	20	20
Optimizer	Adam (without weight decay)		

Table 1: Architectural and training details for each of the three datasets mentioned in the main text.

Architectural details. We tried to use simple architectures throughout our experiments for the sake of interpretability and to emphasize the algorithmic effects as much as possible. For the disks dataset, we used a two layer ConvNet as the inference network, with kernel size 3 and stride 2. The inference networks used for the disks and CelebA-64 datasets have no biases, following Mohan et al. (2022), and the denoisers had no attention blocks. Despite the deliberate restrictions in architecture, we saw that the model was able to learn a robust representation of the dataset. The Half-UNet architecture is a UNet where the Up blocks have been replaced by two MLP layers, both outputting a vector of dimension equal to the latent dimension. From one MLP we get the latent mean. The other outputs a vector which we apply Softplus and squaring operators; this returns the variance.

For the CelebA-HQ 256 dataset, we used an off the shelf DDPM Ho et al. (2020). We froze the denoiser's weights and trained only the inference network, which we again restricted to be a Half-UNet without attention blocks.

KL weight. We annealed the KL weight to the value given in Table 1, starting from a number many orders of magnitude smaller and using a exponential annealing schedule.

Choosing hyperparameters. We performed an extensive grid search over the space of hyperparameters for each model, measuring model performance by the change in MSE between the MSE of an unconditional diffusion model and the MSE when the inference network was also employed.

The unconditional diffusion model retained the same denoiser hyperparameters as the full DiVA. As mentioned in the main text, an informative representation should result in generated images that resemble the original guiding image, and thus a smaller MSE can be used as a measure of latent informativeness. We found that larger KL weight corresponded to more factorized representations, but the weight could only be increased up to a threshold, above which the model exhibits posterior collapse and the representation becomes uninformative. As such, we performed binary search over the KL weights until we found a value just smaller than the threshold.

Timestep sampling distribution. When training the model on the disks and CelebA 64 datasets, we sampled the noise level uniformly from 0 to T, as is common when training unconditional DDPMs. When training the inference network on top of the frozen denoiser on the CelebA-HQ dataset, we found that oversampling the larger noise levels was more effective. This is equivalent to over-weighing the contribution of high noise level terms in the overall MSE loss rather than weighing all levels uniformly. If the inference network is unable to learn the guidance vector for all noise levels when paired with a frozen denoiser, since the guidance vector has greater relative influence at high noise levels compared to the denoising score, it is more effective to learn the guidance score for the high noise regime.

Bottleneck size. We found that the number of semantic attributes learned by the DiVA representation depended on three factors: the magnitude of the KL weight, the dimensionality of the latent representation, and the timestep sampling distribution. Relaxing the first two of these three factors increases the amount of information that can be learned by the network, but results in a less factorized and interpretable representation.

A.6 Comparisons to other models

Property	**DiVA** (ours)	**DiffAE** Preechakul et al. (2022)	**DRL** Mittal et al. (2023)	**InfoDiff** Wang et al. (2023)	**PDAE** Zhang et al. (2022)	**SODA** Hudson et al. (2024)	**DisDiff** Yang et al. (2023)	**DBAE** Kim et al. (2024)
Objective function	ELBO	MSE	MSE+L1 reg.	ELBO+MI reg.	ELBO	MSE	MSE+CE reg.	MSE
Guidance mechanism	Sum	Network†	Network†	Network†	Network†	Network†	Network†	Network†
Probabilistic latents	Yes	No	No	Yes	No	No	No	No
Unsupervised learning	Yes	Yes	Yes	Yes	Yes	No*	Yes	Yes
Using pre-trained DM	Both	No	No	No	Yes	No	No	No
Encodes noisy images	Yes	No	No	No	No	No	No	No
Enc. blind to noise level	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Exact guidance score calc.	Yes	No	No	No	No	No	No	No
Interpretable latent axes	Yes	No	No	Yes	No	No	No	No

Table 2: Comparison to other diffusion-based representation learning models in the literature.

In Table 2 we compare our model to the other diffusion-based representation learning models in the literature. Models with a † sign indicates that they use a separate neural network to perform guidance by approximating the gradient of the log operator. In all networks other than ours, a linear projection of the latent is used to modulate the UNet denoiser's Adaptive GroupNorm layers. The asterisk (*) in the SODA column indicates that while SODA is trained in a self-supervised manner using multiple views of the original datum, their formulation allows for unsupervised training as well. DisDiff uses additional cross entropy losses to encourage representational invariance to a subset of features. However, we find that a combinatorial representation emerges automatically from training our diffusion process to use the exact log likelihood for guidance.

A.7 METRICS

For disentanglement, we measured the degree to which DiVA's representation was disentangled when trained on CelebA 64×64. We used the Total AUROC Difference (TAD) metric, which measures the degree to which a unique attribute is detected by a unique latent representation, and the degree to which the latent axes confidently replicate the ground truth axes (Yeats et al., 2022). Among the diffusion-based representation learning methods, InfoDiff, DisDiff and DBAE are trained to exhibit disentangled representations (Wang et al., 2023; Yang et al., 2023; Kim et al., 2024), so direct comparison against these methods are particularly interesting. We also compared our method against more standard benchmarks in the disentanglement literature, such as VAE and β-VAE.

In Table 3, we see that though we performed little to no hyperparameter search, DiVA performs the best or second best at both disentangling and sample quality metrics. This is likely due to the effect of training our model to produce the exact log likelihood, which requires both clean and noisy images, rather than a representation of only clean images. The coding of noisy images adds important structure to the latent space by encouraging noisy images to have the same latent value as clean images in latent dimensions that code for coarse features. This ensures that coarser features have more global coherence, a hypothesis that is borne out in Fig. 3F and G.

Model	TAD↑	Capt. attrs.↑	FID↓
AE	0.042 ± 0.004	1.0 ± 0.0	90.4 ± 1.8
VAE	0.000 ± 0.000	0.0 ± 0.0	94.3 ± 2.8
β-VAE	0.088 ± 0.051	1.6 ± 0.8	99.8 ± 2.4
InfoVAE	0.000 ± 0.000	0.0 ± 0.0	77.8 ± 1.6
DiffAE (Preechakul et al., 2022)	0.155 ± 0.010	2.0 ± 0.0	22.7 ± 2.1
InfoDiffusion (Wang et al., 2023)	0.299 ± 0.006	3.0 ± 0.0	22.3 ± 1.2
DisDiff (Yang et al., 2023)	0.305 ± 0.010	-	18.2 ± 2.1
DBAE + TC (Kim et al., 2024)	0.362 ± 0.036	3.8 ± 0.8	13.4 ± 0.2
DiVA	0.583	3.0	16.3

Table 3: Comparisons against other generative models on TAD disentanglement metric and FID scores on 64×64 CelebA. "Capt. attrs." refers to the number of captured attributes when calculating the TAD score. Dark gray cells indicate the best, while light gray cells indicate second best.

A.8 DERIVATION FOR GLOBALLY PRESERVED SEMANTIC AXES

We investigated the locality of the mapping between the axes and the semantic label by perturbing the latents of two different images along the same axis. We observed that though in general, this mapping was not preserved across different regions of the latent space, some axes resulted in relatively similar semantic transformations irrespective of the latent.

To identify latent-semantic mappings that are relatively global, we derived a mathematical formulation that provides sufficient conditions for a mapping to be global. Suppose we have a trained conditional diffusion process that maps noisy images $\mathbf{x}_t$ to a clean image estimate $\hat{\mathbf{x}}_0$ via the use of a set of learned latents $\mathbf{z}$. This mapping is defined as a function $f_{\theta,\phi}$ that takes in $\mathbf{z}$ and $\mathbf{x}_t$ and returns a guided clean image estimate:

$$\begin{aligned} f_{\theta,\phi}(\mathbf{x}_t, \mathbf{z}) &= \mathbb{E}_{p_\theta(\mathbf{x}_0|\mathbf{x}_t,\, \mathbf{z})}\big[\mathbf{x}_0\big] \\ &= \mathbb{E}_{p_\theta(\mathbf{x}_0|\mathbf{x}_t)}\big[\mathbf{x}_0\big] + \frac{1-\bar{\alpha}_t}{\sqrt{\bar{\alpha}_t}} \nabla_{\mathbf{x}_t} \log q_\phi(\mathbf{z}|\mathbf{x}_t) \\ &= \mathbf{x}_0^*(\mathbf{x}_t;\theta) + \gamma_t\, \nabla_{\mathbf{x}_t} \log q_\phi(\mathbf{z}|\mathbf{x}_t) \end{aligned} \tag{12}$$

where the analytic form of $\nabla_{\mathbf{x}_t} \log q_\phi(\mathbf{z}|\mathbf{x}_t)$ is given as the derivative of Eq. 11. Given an initial image $\mathbf{x}_0^* = \mathbf{x}_0^*(\mathbf{x}_t, \mathbf{z}')$ that was generated by latent $\mathbf{z}'$, we define the change in the generated image in response to a perturbation to the latent along axis i as a nonlinear function $\psi(\cdot)$ that we can define, without loss of generality, as

$$\mathbf{x}_0^* + \psi(\mathbf{x}_0^*,\ \mathbf{z},\ i) = f(\mathbf{x}_t,\ \mathbf{z} + \alpha\,\hat{\mathbf{i}})$$

where $\alpha\,\hat{\mathbf{i}}$ is a perturbation of magnitude α along a particular axis i. Our aim is to understand whether the functional form of the transformation ψ is dependent on the region of latent space that we are perturbing. In other words, if this image transformation ψ is global, it should be independent of the latent value $\mathbf{z}$, while if the transformation is more local, then the form of ψ should itself change as our test latent point $\mathbf{z}$ changes.

To make this problem tractable, let us assume we are currently in a state with large noise $\mathbf{x}_t$, where $t \approx T$. At high noise levels, the optimal estimate of the clean image that minimizes the mean squared error is the mean of the training set. This suggests that for a well-trained denoiser, we should expect the same output irrespective of the network's input:

$$\hat{\mathbf{x}}_0(\mathbf{x}_t;\theta) = \mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_t)}\big[\mathbf{x}_0\big] \approx \mathbb{E}_{p_{\text{dataset}}(\mathbf{x}_0)}\big[\mathbf{x}_0\big] = \mathbf{c}.$$

where $\mathbf{c}$ is a constant image vector. Under this assumption, our function f is no longer dependent on the denoiser but only depends on the inference network:

$$f_\phi(\mathbf{x}_t,\ \mathbf{z}) = \mathbf{c} + \gamma_t\ \nabla_{\mathbf{x}_t} \log q_\phi(\mathbf{z}|\mathbf{x}_t).$$

This means the transformation ψ is also independent of the denoiser. It is purely a function of the guidance vector:

$$\begin{aligned}\mathbf{x}_0^* + \psi(\mathbf{x}_0^*,\ \mathbf{z},\ i) &= \mathbf{c} + \gamma_t \nabla_{\mathbf{x}_t} \log q_\phi(\mathbf{z} + \alpha\hat{\mathbf{i}}|\mathbf{x}_t) \\ &= \mathbf{c} - \frac{\gamma_t}{2}\nabla_{\mathbf{x}_t}\Big[\big(\mathbf{z} + \alpha\,\hat{\mathbf{i}} - \boldsymbol{\mu}_\phi\big)^\top \Sigma_\phi^{-1}\big(\mathbf{z} + \alpha\,\hat{\mathbf{i}} - \boldsymbol{\mu}_\phi\big) + \log\big|\Sigma_\phi\big| - N\log(2\pi)\Big],\end{aligned}$$

where $\mathbf{z}$ is a function of $\mathbf{x}_0$, and the means $\boldsymbol{\mu}_\phi$ and covariances Σ_ϕ are functions of the current noisy image $\mathbf{x}_t$. If we decompose the RHS of the above equation into the contribution of just $\mathbf{z}$ and the contribution of $\alpha\,\hat{\mathbf{i}}$, and denote the last two terms as the scalar s, we get:

$$\begin{aligned}&\hat{\mathbf{x}}_0^* + \psi(\hat{\mathbf{x}}_0^*,\ \mathbf{z},\ i) \\ &= \mathbf{c} - \frac{\gamma_t}{2}\nabla_{\mathbf{x}_t}\Big[\big(\mathbf{z} - \boldsymbol{\mu}_\phi\big)^\top \Sigma_\phi^{-1}\big(\mathbf{z} - \boldsymbol{\mu}_\phi\big) + 2\big(\mathbf{z} - \boldsymbol{\mu}_\phi\big)^\top \Sigma_\phi^{-1}(\alpha\,\hat{\mathbf{i}}) + (\alpha\,\hat{\mathbf{i}})^\top \Sigma_\phi^{-1}(\alpha\,\hat{\mathbf{i}}) + s\ \Big] \\ &= \mathbf{c} - \frac{\gamma_t}{2}\nabla_{\mathbf{x}_t}\Big[\big(\mathbf{z} - \boldsymbol{\mu}_\phi\big)^\top \Sigma_\phi^{-1}\big(\mathbf{z} - \boldsymbol{\mu}_\phi\big) + 2\alpha\big(\mathbf{z} - \boldsymbol{\mu}_\phi\big)^\top (\Sigma_\phi^{-1})_i + \alpha^2\,(\Sigma_\phi^{-1})_{ii} + s\ \Big],\end{aligned}$$

where $(\Sigma_\phi^{-1})_i$ is the i-th column of the inverse covariance matrix, and $(\Sigma_\phi^{-1})_{ii} = \sigma_i^{-2}$ is the i-th element along the diagonal of the inverse covariance matrix. We can now substitute $\hat{\mathbf{x}}_0^* = \mathbf{c} - \frac{\gamma_t}{2}\nabla_{\mathbf{x}_t}\Big[\big(\mathbf{z} - \boldsymbol{\mu}_\phi\big)^\top \Sigma_\phi^{-1}\big(\mathbf{z} - \boldsymbol{\mu}_\phi\big) + s\ ]$, which yields

$$\psi(\mathbf{z},\ i) = \nabla_{\mathbf{x}_t}\Big[2\alpha(\mathbf{z} - \boldsymbol{\mu}_\phi)^\top (\Sigma_\phi^{-1})_i + \alpha^2\,(\Sigma_\phi^{-1})_{ii}\Big].$$

Let's see how this transformation in $\mathcal{X}$ changes with our location $\mathbf{z}$ in latent space $\mathcal{Z}$. First suppose we use the same latent perturbation $\alpha\,\hat{\mathbf{z}}_i$ irrespective of where we are in latent space. For a given state $\mathbf{x}_t$, we compare two instances of ψ at different values of $\mathbf{z}$:

$$\begin{aligned}\psi_1(\mathbf{z}^{(1)},\ i) &= \nabla_{\mathbf{x}_t}\Big[2\alpha(\mathbf{z}^{(1)} - \boldsymbol{\mu}_\phi)^\top (\Sigma_\phi^{-1})_i + \alpha^2\,(\Sigma_\phi^{-1})_{ii}\Big] \\ \psi_2(\mathbf{z}^{(2)},\ i) &= \nabla_{\mathbf{x}_t}\Big[2\alpha(\mathbf{z}^{(2)} - \boldsymbol{\mu}_\phi)^\top (\Sigma_\phi^{-1})_i + \alpha^2\,(\Sigma_\phi^{-1})_{ii}\Big].\end{aligned}$$

Since $\boldsymbol{\mu}_\phi$ and Σ_ϕ are functions of a common $\mathbf{x}_t$, their values are the same in both ψ_1 and ψ_2. The difference between these two transformations is therefore

$$\begin{aligned}\psi_1(\mathbf{z}^{(1)}) - \psi_2(\mathbf{z}^{(2)}) &= \nabla_{\mathbf{x}_t}\Big[2\alpha(\mathbf{z}^{(1)} - \boldsymbol{\mu}_\phi)^\top (\Sigma_\phi^{-1})_i - 2\alpha(\mathbf{z}^{(2)} - \boldsymbol{\mu}_\phi)^\top (\Sigma_\phi^{-1})_i\Big] \\ &= 2\alpha\nabla_{\mathbf{x}_t}\Big[(\mathbf{z}^{(1)} - \mathbf{z}^{(2)})^\top (\Sigma_\phi^{-1})_i\Big].\end{aligned}$$

We see that the difference in the transformation scales linearly with the difference in $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$, and scales with the i-th column of the inverse covariance matrix. If we assume a diagonal covariance assumption, as in our model, this is simply a unit vector scaled by the i-th element along the diagonal:

$$\begin{aligned}\psi_1(\mathbf{z}^{(1)}) - \psi_2(\mathbf{z}^{(2)}) &= 2\alpha\,\nabla_{\mathbf{x}_t}\Big[(\mathbf{z}^{(1)} - \mathbf{z}^{(2)})^\top \sigma_i^{-2}\hat{\mathbf{z}}_i\Big] \\ &= 2\alpha\,\big(\nabla_{\mathbf{x}_t}\sigma_i^{-2}\big)\big(\mathbf{z}_i^{(1)} - \mathbf{z}_i^{(2)}\big),\end{aligned} \tag{13}$$

where in the second line we use the fact that only σ_i^{-2} is a function of $\mathbf{x}_t$. This means that if the clean images are mapped to similar values in this dimension (e.g. if the population variance in this dimension is very low), then the function that is applied in the data space will be very similar. The effect is also scaled by the derivative of the precision: we want to look for axes where the precision isn't affected by perturbations around $\mathbf{x}_t$:

$$\nabla_{\mathbf{x}_t}\sigma_i^{-2}(\mathbf{x}_t) = \frac{\partial}{\partial\sigma_i^2}\left(\frac{1}{\sigma_i^2}\right)\frac{\partial\sigma_i^2}{\partial\mathbf{x}_t} = -\frac{1}{\sigma_i^4}\frac{\partial\sigma_i^2}{\partial\mathbf{x}_t}.$$

Plugging this into Eq. 13 yields

$$\psi_1(\mathbf{z}^{(1)}) - \psi_2(\mathbf{z}^{(2)}) = -\frac{2\alpha}{\sigma_i^4}\frac{\partial\sigma_i^2}{\partial\mathbf{x}_t}\left(\mathbf{z}_i^{(1)} - \mathbf{z}_i^{(2)}\right).$$

Global axes are those that minimize the difference in ψ. We can find axes that satisfy three properties:

- small variance in the population of clean image latents,
- large feature uncertainty σ_i for our noisy image state $\mathbf{x}_t$,
- where the variance of the noisy latent σ_i is insensitive to changes in our noisy image $\mathbf{x}_t$.

Though our analysis relies on the assumption that we are currently at a large noise regime, this is also the regime in which the guidance vector has the most influence over the sampling process, since the noise-dependent factor γ_t in Eq. 12 —which trades off between the influence of the guidance vector and the denoiser score—scales monotonically with t.

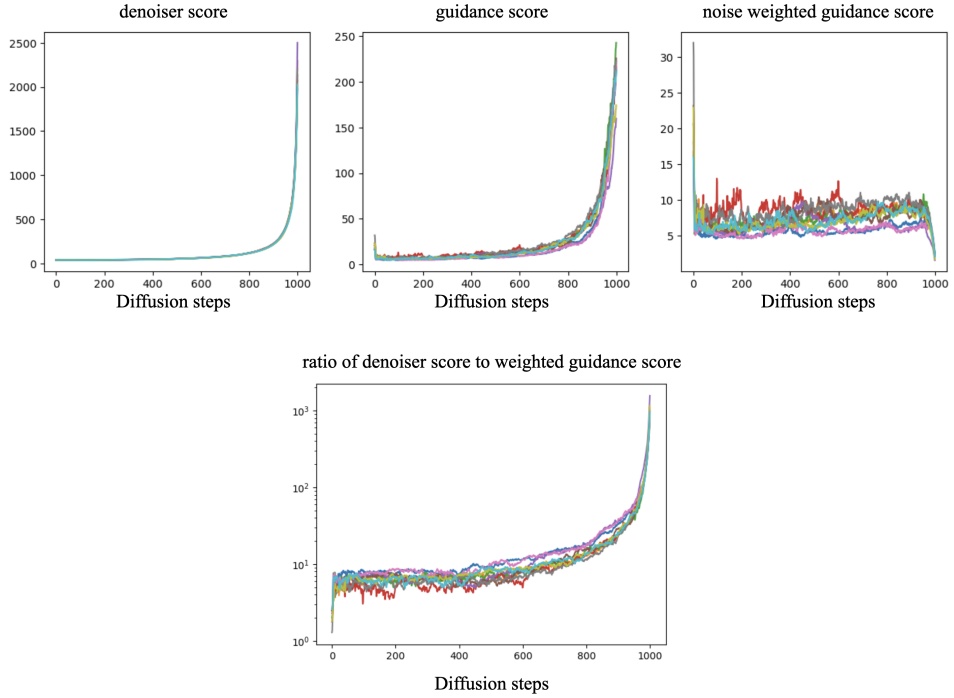

Figure 5: Comparing the norms of the denoiser score and the guidance score. Larger (rightward) denoising steps corresponds to smaller noise level. The ratio is closest to one at the start of the diffusion process, at large noise levels.

Above, we plot the ratio between the norm of the denoiser score and the norm of the weighted guidance score. A smaller ratio indicates that the guidance score has more relative strength. We see that the ratio is smallest at large noise levels, at the start of the diffusion process.