

---

# Towards Understanding the Mechanisms of Classifier-Free Guidance

---

Xiang Li<sup>1</sup>, Rongrong Wang<sup>2</sup>, Qing Qu<sup>1</sup>

<sup>1</sup>Department of EECS, University of Michigan,

<sup>2</sup>Department of CMSE and Mathematics, Michigan State University,

[forkobe@umich.edu](mailto:forkobe@umich.edu), [wangron6@msu.edu](mailto:wangron6@msu.edu), [qingqu@umich.edu](mailto:qingqu@umich.edu)

## Abstract

Classifier-free guidance (CFG) is a core technique powering state-of-the-art image generation systems, yet its underlying mechanisms remain poorly understood. In this work, we begin by analyzing CFG in a simplified linear diffusion model, where we show its behavior closely resembles that observed in the nonlinear case. Our analysis reveals that linear CFG improves generation quality via three distinct components: (i) a mean-shift term that approximately steers samples in the direction of class means, (ii) a positive Contrastive Principal Components (CPC) term that amplifies class-specific features, and (iii) a negative CPC term that suppresses generic features prevalent in unconditional data. We then verify these insights in real-world, *nonlinear* diffusion models: over a broad range of noise levels, linear CFG resembles the behavior of its nonlinear counterpart. Although the two eventually diverge at low noise levels, we discuss how the insights from the linear analysis still shed light on the CFG’s mechanism in the nonlinear regime.

## 1 Introduction

Diffusion models [1–4] generate samples from a data distribution  $p_{\text{data}}(\mathbf{x})$ , where  $\mathbf{x} \in \mathbb{R}^d$ , by reversing a forward noising process. This forward process, defined in (1), progressively corrupts the clean data until  $p(\mathbf{x}; \sigma_{\max})$  becomes indistinguishable from a Gaussian distribution  $\mathcal{N}(\mathbf{0}, \sigma_{\max}^2 \mathbf{I})$ ,

$$p(\mathbf{x}; \sigma(t)) = \int_{\mathbb{R}^d} p_{0t}(\mathbf{x}|\mathbf{x}_0) p_{\text{data}}(\mathbf{x}_0) d\mathbf{x}_0. \quad (1)$$

Following the state-of-the-art EDM framework [4, 5], the forward transition kernel is set to  $p_{0t}(\mathbf{x}|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0, \sigma^2(t)\mathbf{I})$ . The reverse process can then be expressed as a probabilistic ODE:

$$d\mathbf{x}_t = -\sigma(t) \nabla_{\mathbf{x}_t} \log p(\mathbf{x}; \sigma(t)) dt, \quad (2)$$

such that  $\mathbf{x}_t \sim p(\mathbf{x}; \sigma(t))$  for every  $\sigma(t) \in (0, \sigma_{\max}]$ . In practice, the score function can be approximated as  $\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma(t)) \approx (\mathcal{D}_{\theta}(\mathbf{x}; \sigma(t)) - \mathbf{x})/\sigma^2(t)$ , where  $\mathcal{D}_{\theta}$  is a deep network-based denoiser with parameter  $\theta$  optimized by minimizing the denoising score matching objective [6]:

$$\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2(t)\mathbf{I})} [||\mathcal{D}_{\theta}(\mathbf{x} + \epsilon; \sigma(t)) - \mathbf{x}||_2^2]. \quad (3)$$

To sample from conditional distribution  $p(\mathbf{x}|\mathbf{c})$ , the deep denoiser  $\mathcal{D}_{\theta}(\mathbf{x}; \sigma(t), \mathbf{c})$  receives an auxiliary embedding  $\mathbf{c}$  specifying the target class or other conditions during training such that conditional sampling can be performed with:

$$d\mathbf{x}_t = -\sigma(t) \nabla_{\mathbf{x}_t} \log p(\mathbf{x}|\mathbf{c}; \sigma(t)) dt, \quad (4)$$

where  $\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{c}; \sigma(t)) \approx (\mathcal{D}_{\theta}(\mathbf{x}; \sigma(t), \mathbf{c}) - \mathbf{x})/\sigma^2(t)$ . However, the naive (standard) conditional sampling (4) alone often results in images with incoherent structures and fail to align well with the

target condition [7]. Classifier-free guidance (CFG) [8] addresses this issue by steering the naive conditional sampling trajectory with a *guidance term*:

$$g(\mathbf{x}, t) = \nabla_{\mathbf{x}} \log p(\mathbf{x} | \mathbf{c}; \sigma(t)) - \nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma(t)), \quad (5)$$

so that (4) becomes:

$$d\mathbf{x}_t = -\sigma(t)(\nabla_{\mathbf{x}_t} \log p(\mathbf{x} | \mathbf{c}; \sigma(t)) + \gamma g(\mathbf{x}, t))dt, \quad (6)$$

where  $\gamma \geq 0$  controls the strength of guidance. With a properly chosen  $\gamma$ , CFG substantially improves sample quality, albeit with reduced diversity. Since its invention, CFG and its variants [9–15] have become the backbone that powers the most advanced image generation systems [16–18].

Despite the practical success of CFG, its underlying mechanism remains largely unknown. As shown in [7], the CFG-perturbed reverse trajectory does not correspond to any known forward process, therefore, analyzing the effects of CFG requires case-by-case studies with explicit assumptions on the data distribution. For example, work [19] proves that under an isotropic Gaussian mixture data assumption, CFG boosts classification accuracy at the cost of sample diversity. The work [20] shows that under either 1-D mixtures of compactly supported distributions or 1-D isotropic Gaussian data assumptions, CFG guides the diffusion models towards sampling more heavily from the boundary of the support. Despite providing invaluable insights, these analyses rely on oversimplified assumptions that neglect critical aspects of real data, particularly the covariance structures of natural images. Consequently, it remains unclear how well these theoretical results generalize to diffusion models trained on complex image datasets.

In this work, we pursue a deeper understanding CFG’s mechanism, focusing on two core questions: (i) *What is the failure mode of naive conditional sampling, i.e., in what aspect is the generated images subpar compared to the training images?* and (ii) *how does CFG mitigate this problem?*

To answer the first question, we show that the naive conditional suffers from a lack of class-specificity issue: images conditioned on different labels often share similar structures and lack distinct class features. We posit that this issue can be partially attributed to the covariance structures of different classes being insufficiently distinct. Recent studies [21, 22] observe that over a broad range of noise levels, diffusion models can be *unreasonably* approximated by the optimal linear denoisers for the multivariate Gaussian distribution defined by the empirical mean and covariance of the training set. Consequently, the data covariance (and particularly its principal components, or PCs) heavily influences the generation. However, as we will demonstrate, different classes can share overly similar covariance structures, resulting in generated images that lack class-specific patterns.

Based on this intuition, we posit that CFG must identify the *unique* features of the target class. To understand how this is achieved, we study the prototypical setting of the optimal *linear* diffusion model, where we show that CFG guidance naturally decomposes into three components with distinct effects: (i) a *mean-shift* term that approximately pushes the samples towards the direction of the class mean, (ii) a *positive contrastive principal components (CPC)* term that enhances the target class’s unique features and (iii) a *negative contrastive principal components (CPC)* term that suppresses the features prominent in the unconditional dataset. Despite the simplicity of the linear model, the linear CFG greatly improves the visual quality of generated samples in a way reminiscent of real-world, nonlinear deep diffusion models, implying that nonlinear CFG share a similar underlying working mechanism. We then investigate how well the insights derived from the linear setting extend to actual diffusion models. We first show that at high to moderate noise levels, linear CFG yields highly similar effects as those of the nonlinear CFG. As noise decreases further and the diffusion model enters a highly nonlinear regime, the effects of linear CFG and actual nonlinear CFG begin to diverge. Nevertheless, by interpreting denoising as weighted projection onto an adaptive basis, the insights from linear analysis can still shed light on the CFG’s mechanism in the nonlinear regime.

**Contributions.** Our main contributions are as follows:

- We identify the lack of class-specificity issue of naive conditional sampling, linking it to the non-distinctiveness of class covariances. Under a linear model assumption, we show CFG overcomes this issue by amplifying class-specific features, suppressing unconditional ones and shifting the samples in the direction of class mean.
- We validate these insights derived in the linear model on real diffusion models, demonstrating that: (i) at high to moderate noise levels, linear CFG closely matches the effects of nonlinear CFG, and (ii) at low noise levels, the insights from the linear analysis can still shed light on the mechanism of CFG in this nonlinear regime.

## 2 Preliminaries

### 2.1 Optimal Linear Diffusion Model

Suppose  $p_{\text{data}}(\mathbf{x})$  has mean  $\boldsymbol{\mu}$  and covariance  $\boldsymbol{\Sigma}$ . Under the constraint that  $\mathcal{D}(\mathbf{x}; \sigma(t))$  is a linear model (with a bias term), the optimal solution to (3) has the analytical form:

$$\mathcal{D}_L(\mathbf{x}; \sigma(t)) = \boldsymbol{\mu} + \mathbf{U} \tilde{\boldsymbol{\Lambda}}_{\sigma(t)} \mathbf{U}^T (\mathbf{x} - \boldsymbol{\mu}), \quad (7)$$

where  $\boldsymbol{\Sigma} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^T$  is the full SVD of the covariance matrix,  $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_d)$  is the singular values and  $\tilde{\boldsymbol{\Lambda}}_{\sigma(t)} = \text{diag}\left(\frac{\lambda_1}{\lambda_1 + \sigma^2(t)}, \dots, \frac{\lambda_d}{\lambda_d + \sigma^2(t)}\right)$ . With this linear denoiser, the reverse diffusion ODE (2) has the following closed-form expression (see appendix B.1 for the proof):

$$\mathbf{x}_t = \boldsymbol{\mu} + \sum_{i=1}^d \sqrt{\frac{\lambda_i + \sigma^2(t)}{\lambda_i + \sigma^2(T)}} \mathbf{u}_i^T (\mathbf{x}_T - \boldsymbol{\mu}) \mathbf{u}_i, \quad (8)$$

where  $T$  is the starting timestep and  $\mathbf{u}_i$  is the  $i^{\text{th}}$  singular vector of  $\boldsymbol{\Sigma}$ , which is also the  $i^{\text{th}}$  principal component. Note that in this linear setting, the generated samples are largely determined by the data covariance.

Recent studies [21, 22] show that for a wide range (high to moderate) of noise levels, deep network-based diffusion models can be well approximated by the linear model (7), with  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  set to the empirical mean and covariance of the training data. As shown in Figures 1 and 14, the sampling trajectories of the deep diffusion model (EDM) and the linear model share high similarity at high to moderate noise levels. Although the models begin to diverge at lower noise levels—where EDM exhibits strong nonlinearity and realistic image content begins to form—their final samples still share a similar overall structure. Moreover, it is shown in [21] that this similarity is particularly obvious when the deep network has limited capacity or the training is insufficient. Since  $\mathcal{D}_L(\mathbf{x}; \sigma(t))$  is the optimal denoiser for  $p(\mathbf{x}; \sigma(t))$  induced by  $p_{\text{data}}(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , sampling with  $\mathcal{D}_L$  is equivalent to sampling from  $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Hence, we refer to  $\mathcal{D}_L$  as the *linear Gaussian model*.

### 2.2 Contrastive Principal Component Analysis

Principal component analysis (PCA) [23, 24] identifies directions that capture the most variances in a dataset. These principal components (PCs), which are equivalent to the singular vectors of the data covariance matrix, are widely used for data exploration and visualization. However, large variance alone does not guarantee that a PC captures the unique patterns tied to the dataset; it may instead reflect more general patterns such as foreground-background variations.

To discover low-dimensional structure that is unique to a dataset, the work [25] proposed the contrastive principal component analysis (CPCA), which utilizes a *background* (or reference) dataset to highlight patterns unique to the *target* dataset. Let  $X$  and  $Y$  be two datasets with covariance matrices  $\boldsymbol{\Sigma}_X$  and  $\boldsymbol{\Sigma}_Y$ , respectively. For a unit vector  $\mathbf{v} \in \mathbb{S}^{d-1}$ , its variances  $\text{Var}_X(\mathbf{v})$  and  $\text{Var}_Y(\mathbf{v})$  in the two datasets are:

$$\text{Var}_X(\mathbf{v}) := \mathbf{v}^T \boldsymbol{\Sigma}_X \mathbf{v}, \quad \text{Var}_Y(\mathbf{v}) := \mathbf{v}^T \boldsymbol{\Sigma}_Y \mathbf{v}. \quad (9)$$

If  $\mathbf{v}$  corresponds to a unique class-specific pattern of  $X$ , we expect  $\text{Var}_X(\mathbf{v}) \gg \text{Var}_Y(\mathbf{v})$ , i.e., it explains significantly more variance in  $X$  than in  $Y$ . Such directions, called the *contrastive principal components (CPCs)*, can be found by maximizing:

$$\arg \max_{\mathbf{v} \in \mathbb{S}^{d-1}} \mathbf{v}^T (\boldsymbol{\Sigma}_X - \boldsymbol{\Sigma}_Y) \mathbf{v}, \quad (10)$$

which are essentially the top eigenvectors of  $\boldsymbol{\Sigma}_X - \boldsymbol{\Sigma}_Y$ . Geometrically, the first  $k$  CPCs span the  $k$ -dimensional subspace that best fits the dataset  $X$  while being as far as possible from  $Y$  (see appendix A for details). Conversely, directions  $\mathbf{v}$  for which  $\text{Var}_X(\mathbf{v}) \approx \text{Var}_Y(\mathbf{v})$  represent either universal structures shared by both  $X$  and  $Y$  or meaningless features lying in the null space of the data covariances—and are thus discarded as less interesting. Finally, a scalar factor can be introduced in (10) to control the strength of the contrast.

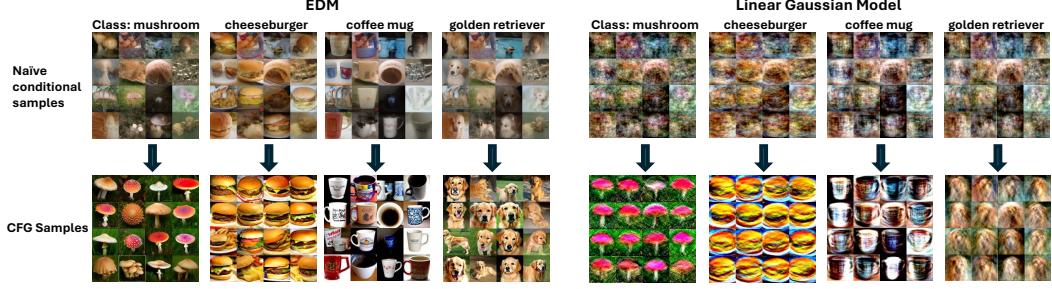


Figure 2: **Effects of CFG.** Left and right figures compare the naive conditional sampling (top rows) versus CFG-guided sampling (bottom rows) for deep diffusion models (EDM) and linear Gaussian diffusion models, respectively. Each grid cell corresponds to the same initial noise. While naive conditional samples lack class-specific clarity, CFG significantly improves both visual quality and distinctiveness. The conditional linear models are built with class-specific means and covariances. Please refer to appendix D for more experiment results.

### 2.3 Posterior Data Covariance

Consider  $\mathbf{x} \sim p_{\text{data}}(\mathbf{x})$  and  $\mathbf{x}_t = \mathbf{x} + \sigma(t)\epsilon$ , where  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . Then the posterior covariance of  $p(\mathbf{x}|\mathbf{x}_t)$ , denoted by  $\text{Cov}[\mathbf{x}|\mathbf{x}_t]$ , is proportional to the denoiser's Jacobian [26]:

$$\text{Cov}[\mathbf{x}|\mathbf{x}_t] = \sigma^2(t) \nabla \mathcal{D}(\mathbf{x}_t; \sigma(t)), \quad (11)$$

where  $\nabla \mathcal{D}(\mathbf{x}_t; \sigma(t)) = \frac{\partial \mathcal{D}(\mathbf{x}_t; \sigma(t))}{\partial \mathbf{x}_t}$  is the Jacobian of the optimal denoiser  $\mathcal{D}(\mathbf{x}; \sigma(t))$  at input  $\mathbf{x}_t$ . Analogous to PCs, the singular vectors of  $\text{Cov}[\mathbf{x}|\mathbf{x}_t]$  are the *posterior PCs*, representing directions of maximal variances of all clean images that could have generated the noisy observation  $\mathbf{x}_t$ . In the case that  $p_{\text{data}} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , we have  $\text{Cov}[\mathbf{x}|\mathbf{x}_t] = \sigma^2(t) \mathbf{U} \tilde{\boldsymbol{\Lambda}}_{\sigma(t)} \mathbf{U}^T$ , matching  $\nabla \mathcal{D}_{\text{L}}(\mathbf{x}_t; \sigma(t))$ , the Jacobians of the optimal linear denoiser (7), and is independent of  $\mathbf{x}_t$ . In more general scenarios, one can approximate  $\text{Cov}[\mathbf{x}|\mathbf{x}_t]$  by computing the network Jacobian at  $\mathbf{x}_t$  via automatic differentiation.

## 3 Analyzing CFG in Linear Model

In this section, we first show that naive conditional sampling often produces low-quality samples lacking clear class-specific features, which we attribute to the non-distinctiveness of class covariance matrices (section 3.1). We then theoretically analyze how CFG in the context of linear diffusion models alleviates this issue (section 3.2).

### 3.1 Naive Conditional Generation Lacks Class-Specificity

Figure 2(left) (top row) shows samples generated via naive conditional sampling (4). Qualitatively, these samples often exhibit poor visual fidelity, with incoherent features that blend into the background and the class-specific image structures can be hard to discriminate. Moreover, even when conditioned on different class labels, images generated from the same initial noise share high structural similarity, suggesting that naive conditional sampling fails to capture discriminative, class-dependent patterns.

To quantify this loss of class-specificity, we compute the pairwise inter-class similarity with the FID metric [27]. For each pair of classes, we construct two datasets  $X$  and  $Y$  and evaluate the FID between them. As shown in Figure 3, when  $X$  and  $Y$  are built with images generated with naive conditional sampling, the FID (colored in orange) is consistently lower than when they are built with the training data (colored in blue). Since lower FID indicates higher similarity, this result confirms that compared with the training images, which represent the ground truth data distribution, images generated by naive conditional sampling are less distinguishable across classes.

This issue is especially pronounced in linear diffusion models. As shown in Figure 2(right, top row), samples generated with linear diffusion models built with class-specific means and covariances appear

FID between classes: training data / naive conditional samples / CFG-guided samples				
tach		tree frog	green mamba	golden retriever
tach	0	223.8 / 216.0 / 279.1	238.4 / 227.4 / 247.0	240.4 / 223.6 / 271.6
tree frog	223.8 / 216.0 / 279.1	0	167.2 / 113.3 / 254.6	237.0 / 212.9 / 278.5
green mamba	238.4 / 227.4 / 247.0	167.2 / 113.3 / 254.6	0	281.8 / 249.0 / 280.6
golden retriever	240.4 / 223.6 / 271.6	237.0 / 212.9 / 278.5	281.8 / 249.0 / 280.6	0

Figure 3: **Class-to-Class Similarity.** Each cell reports the FID between datasets of two classes, built with (i) training data (ii) data generated by naive conditional sampling and (iii) data generated by CFG sampling (refer to appendix D.2 for experiment details and more results.)

highly similar. From (8), we see that the linear sampling trajectory is governed by the data covariance:  $\mathbf{x}_t$  is a linear combination of PCs, weighted by (i) the correlation  $u_i^T(\mathbf{x}_T - \boldsymbol{\mu})$  between the mean-subtracted initial noise  $\mathbf{x}_T$  and the  $i$ -th PC, and (ii) scaling factors  $\sqrt{\frac{\lambda_i + \sigma^2(t)}{\lambda_i + \sigma^2(T)}}$  that emphasize leading PCs. Consequently, if class-conditional covariances lack sufficiently discriminative structures (which is indeed true as shown in appendix D.3), generated samples will appear similar regardless of class label. This lack of class-specificity aligns with prior findings [25], which shows that PCs often capture generic image variations (e.g., foreground-background), rather than class-specific patterns.

The existence of the class-specificity gap implies these models fail to fully capture the higher-order statistics of the training data: if they did, naive conditional sampling, which by construction samples from the target conditional distribution, would already produce high quality samples, and CFG would only distort the target distribution. Linear diffusion models represent an extreme case: due to the linear constraint, they can only learn the first and second-order moments (mean and covariance) of the training data, which despite being fundamental data statistics, cannot capture the rich, nonlinear dependencies necessary for realistic generation. In particular, when covariances across classes share high similarity, samples initialized from the same noise become visually alike regardless of label.

We hypothesize that real-world diffusion models inherit similar limitations. Although nonlinear diffusion models surely learn beyond second-order statistics, as discussed in section 2.1, for high to moderate noise levels, they can be well approximated by linear models, especially under limited model capacity or insufficient training. Indeed, Figure 2 (top row) and Figure 13, 14 demonstrate that linear models reproduce the coarse-grained structures of nonlinear diffusion samples, implying that the covariance structure plays a significant role in shaping the high-level features of the generated samples. These observations reflect a well-known *simplicity bias*, where deep networks favor learning low-order, linearly structured representations over complex, higher-order dependencies [28]. Hence, if the covariances are indistinct across classes, sample quality can be limited even in nonlinear models (see appendix D.3 for more discussion).

As quantitatively shown in Figure 3, CFG significantly increases the inter-class separation: FID (colored in green) between different generated classes rises. Qualitatively, Figure 2 (bottom row) shows that CFG substantially improves both linear and nonlinear models, producing visibly better samples with enhanced class-specific structures. Similar effects of CFG across both linear and nonlinear models motivate us to use a linear model as a simplified prototype to analyze how CFG reshapes the generation process and why it is effective.

### 3.2 How Linear CFG Leads to Distinct Generations

We now dissect how CFG, in the linear diffusion models, produces samples with distinct class-specific features. Consider two independent optimal linear denoisers,  $\mathcal{D}_L(\mathbf{x}_t; \sigma(t), \mathbf{c})$  for conditional data and  $\mathcal{D}_L(\mathbf{x}_t; \sigma(t))$  for unconditional data, with means  $\boldsymbol{\mu}_c, \boldsymbol{\mu}_{uc}$  and covariances  $\boldsymbol{\Sigma}_c = \mathbf{U}_c \boldsymbol{\Lambda}_c \mathbf{U}_c^T$  and  $\boldsymbol{\Sigma}_{uc} = \mathbf{U}_{uc} \boldsymbol{\Lambda}_{uc} \mathbf{U}_{uc}^T$ , respectively. Substituting the optimal linear denoiser (7) into (6), the CFG-guided sampling process can be decomposed into three terms:

$$d\mathbf{x}_t = -\sigma(t)(f_c(\mathbf{x}_t, t) + g_{cpc}(\mathbf{x}_t, t) + g_{mean}(t))dt, \quad (12)$$

where by letting  $\tilde{\boldsymbol{\Sigma}}_{c,t} = \mathbf{U}_c \tilde{\boldsymbol{\Lambda}}_{\sigma(t),c} \mathbf{U}_c^T$  and  $\tilde{\boldsymbol{\Sigma}}_{uc,t} = \mathbf{U}_{uc} \tilde{\boldsymbol{\Lambda}}_{\sigma(t),uc} \mathbf{U}_{uc}^T$ , each term takes the following form: (i)  $f_c(\mathbf{x}_t, t) = \frac{1}{\sigma^2(t)}(\tilde{\boldsymbol{\Sigma}}_{c,t} - \mathbf{I})(\mathbf{x}_t - \boldsymbol{\mu}_c)$ , (ii)  $g_{cpc}(\mathbf{x}_t, t) = \frac{\gamma}{\sigma^2(t)}(\tilde{\boldsymbol{\Sigma}}_{c,t} - \tilde{\boldsymbol{\Sigma}}_{uc,t})(\mathbf{x}_t - \boldsymbol{\mu}_c)$ , and (iii)  $g_{mean}(t) = \frac{\gamma}{\sigma^2(t)}(\mathbf{I} - \tilde{\boldsymbol{\Sigma}}_{uc,t})(\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc})$ .

Here,  $f_c(\mathbf{x}_t, t)$  is the standard conditional score, and  $g_{cpc}(\mathbf{x}_t, t)$  plus  $g_{mean}(t)$  form the CFG guidance (derivation for the decomposition is provided in appendix B.2). Let  $\mathbf{V}_{\sigma(t)} \hat{\boldsymbol{\Lambda}}_{\sigma(t)} \mathbf{V}_{\sigma(t)}^T$  be the eigen decomposition of  $\tilde{\boldsymbol{\Sigma}}_{c,t} - \tilde{\boldsymbol{\Sigma}}_{uc,t}$ , whose spectrum contains both positive and negative eigenvalues (see Figure 18),  $g_{cpc}(\mathbf{x}_t, t)$  can be split accordingly into *positive* and *negative* CPC components:

$$\underbrace{\frac{\gamma}{\sigma^2(t)}(\mathbf{V}_{\sigma(t),+} \hat{\boldsymbol{\Lambda}}_{\sigma(t),+} \mathbf{V}_{\sigma(t),+}^T)(\mathbf{x}_t - \boldsymbol{\mu}_c)}_{\text{positive CPC guidance}} = \frac{\gamma}{\sigma^2(t)} \sum_i \hat{\lambda}_{+,i} \mathbf{v}_{+,i} (\mathbf{v}_{+,i}^T(\mathbf{x}_t - \boldsymbol{\mu}_c)), \quad (13)$$

$$\underbrace{\frac{\gamma}{\sigma^2(t)}(\mathbf{V}_{\sigma(t),-} \hat{\boldsymbol{\Lambda}}_{\sigma(t),-} \mathbf{V}_{\sigma(t),-}^T)(\mathbf{x}_t - \boldsymbol{\mu}_c)}_{\text{negative CPC guidance}} = \frac{\gamma}{\sigma^2(t)} \sum_i \hat{\lambda}_{-,i} \mathbf{v}_{-,i} (\mathbf{v}_{-,i}^T(\mathbf{x}_t - \boldsymbol{\mu}_c)), \quad (14)$$

where  $\mathbf{V}_{\sigma(t),+}$  and  $\mathbf{V}_{\sigma(t),-}$  contain eigenvectors  $\mathbf{v}_{+,i}$  and  $\mathbf{v}_{-,i}$  corresponding to positive and negative eigenvalues  $\hat{\Lambda}_{\sigma(t),+}$  and  $\hat{\Lambda}_{\sigma(t),-}$  respectively. As discussed in section 2.3,  $\tilde{\Sigma}_{c,t}$  and  $\tilde{\Sigma}_{uc,t}$  are up to a scaling factor  $\sigma^2(t)$  equivalent to the conditional and unconditional posterior covariances of  $p_{\text{data}}(\mathbf{x}|\mathbf{c}) = \mathcal{N}(\boldsymbol{\mu}_c, \Sigma_c)$  and  $p_{\text{data}}(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{uc}, \Sigma_{uc})$ . Hence,  $\mathbf{V}_{\sigma(t)}$  are the CPCs which contrast between  $X \sim p_{\text{data}}(\mathbf{x}|\mathbf{x}_t, \mathbf{c})$  and  $Y \sim p_{\text{data}}(\mathbf{x}|\mathbf{x}_t)$ . Specifically,  $\mathbf{V}_{\sigma(t),+}$  captures directions of higher conditional variance (class-specific features), while  $\mathbf{V}_{\sigma(t),-}$  captures directions of higher unconditional variance (features more prevalent in the unconditional data).

**Distinctive Effects of the CFG Components.** Figure 4(a) shows that for both nonlinear (EDM) and linear models, CFG significantly enhances the characteristic pattern—a person holding a fish—of the "tench" class from ImageNet [29]. Next, we isolate the roles of each CFG term by selectively enabling only one at a time within the linear model. In the following discussion, we omit the negative sign in (12) since the ODE runs backward in time:

- The **positive CPC term** (13) projects  $\mathbf{x}_t - \boldsymbol{\mu}_c$  onto the subspace spanned by the positive CPCs, i.e., the eigenvectors  $\mathbf{v}_{+,i}$  associated with positive eigenvalues, with each component scaled by its eigenvalue  $\hat{\lambda}_{+,i} \geq 0$ , (13) is added to  $\mathbf{x}_t$ , i.e., the components of  $\mathbf{x}_t - \boldsymbol{\mu}_c$  that align with the positive CPCs, which represent the class-specific features, are amplified. Figure 4 (b) (second column) show the first 25 positive CPCs of  $\Sigma_c - \Sigma_{uc}$ <sup>1</sup> and the resulting samples. Compared to the conditional PCs of the dataset, the positive CPCs better capture the unique patterns of the class, which emerge visibly in the generated images.
- Similarly, the **negative CPC term** (14) projects  $\mathbf{x}_t - \boldsymbol{\mu}_c$  onto the negative CPC directions  $\mathbf{v}_{-,i}$ . Since  $\hat{\lambda}_{-,i} < 0$ , these components are subtracted from  $\mathbf{x}_t$ , suppressing features associated more strongly with the unconditional data. Figure 4 (b) (third column) shows the first 25 negative CPCs and the resulting generations. Although visually less interpretable, these directions represent common but target-class-irrelevant features in the unconditional data. Suppressing them reduces background clutter and irrelevant content, making class-relevant structures more salient.
- In the context of linear diffusion model, it can be shown that (see proof in appendix B.2):

$$g_{\text{mean}}(t) = \gamma \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{c}; \sigma(t))} [\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{c}; \sigma(t)) - \nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma(t))]. \quad (15)$$

Thus, the **Mean-shift term**  $g_{\text{mean}}(t)$  can be interpreted as the probability-weighted average of the steepest ascent direction that maximizes the difference (log-likelihood ratio) of the noise-mollified conditional and unconditional distributions. Note that when  $\sigma(t)$  is large, where  $\mathbf{I} - \tilde{\Sigma}_{uc,t} \approx \mathbf{I}$ ,  $g_{\text{mean}}(t)$  approximately shifts  $\mathbf{x}_t$  in the direction of  $\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc}$ , which is the difference between conditional and unconditional mean. As  $\sigma(t)$  decreases,  $\mathbf{I} - \tilde{\Sigma}_{uc,t}$  progressively shrink the components  $\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc}$  lying in the column space of  $\mathbf{U}_{uc}$  (the covariances of image datasets are typically low-rank), while preserving its energy in the null space.

Figure 4(b), fourth column, shows that the mean-shift term enhances the structure of class mean in the generated samples. However, unlike the positive CPC term,  $g_{\text{mean}}(t)$  is independent of  $\mathbf{x}_t$ , thus producing more homogeneous samples with reduced diversity.

**Analytical Solution to the CFG Trajectory.** To better understand the distinct effects of the CFG components, we aim to examine the global solution to the linear ODE system (12). However, the variables in the general solution of (12) are coupled and difficult to interpret. To obtain a more tractable expression, we follow [30, 31] to make the following assumption:

**Assumption 1.** *The covariance matrices  $\Sigma_c$  and  $\Sigma_{uc}$  are simultaneously diagonalizable, i.e.,  $\Sigma_{uc} = \mathbf{U}_c \boldsymbol{\Lambda}_{uc} \mathbf{U}_c^T$ , where  $\mathbf{U}_c \in \mathbb{R}^d$  are the singular vectors (principal components) of the conditional covariance. Here  $\boldsymbol{\Lambda}_{uc}$  is not necessarily ordered by the magnitude of the singular values.*

Assumption 1, known as the *Common Principal Components Assumption* is widely applied to analyze structural relationships across data groups. Under this assumption, the relative importance of the  $i$ -th principal component  $\mathbf{u}_{c,i}$  in the conditional and unconditional datasets is determined by the relative magnitudes of its associated singular values. If  $\lambda_{c,i} > \lambda_{uc,i}$ , then  $\mathbf{u}_{c,i}$  is the positive CPC as it captures more variance in the conditional distribution, while  $\lambda_{c,i} < \lambda_{uc,i}$  implies that  $\mathbf{u}_{c,i}$  is more relevant to the unconditional distribution; therefore, it is a negative CPC.

<sup>1</sup>Although  $\mathbf{V}_{\sigma(t)}$  depends on  $\sigma(t)$ , over a wide range of noise levels (especially high ones), it remains close to the eigenvectors of  $\Sigma_c - \Sigma_{uc}$ . We provide its full evolution across time in Figure 19.

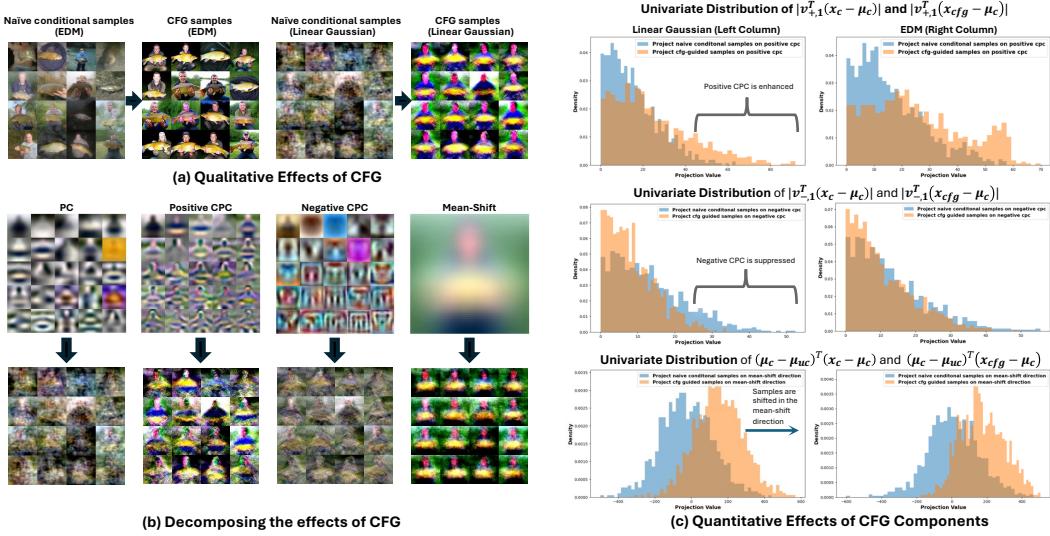


Figure 4: **Distinct effects of different CFG components.** (a) CFG substantially enhances class-specific features (in both EDM and linear diffusion). (b) Top row: PCs, positive/negative CPCs, and  $\mu_c - \mu_{uc}$ . Bottom row: generated samples when each component is applied in isolation. (c) One-dimensional densities of generated samples after projection onto key directions. The left column corresponds to the linear diffusion model, whereas the right column corresponds to the EDM model. Top row: project onto leading positive CPC. Middle row: project onto negative CPC. Third row: project onto the mean-shift direction. Here we only plot the resulting histograms for the first positive and negative CPCs but the same patterns hold for subsequent CPCs. For experimental details and more results, please refer to appendix E.3.

**Theorem 1.** Under Assumption 1, the solution to the linear CFG process (12) is:

$$\mathbf{x}_t = \mu_c + \sum_{i=1}^d h(\lambda_{c,i}, \lambda_{uc,i})^{\frac{\gamma}{2}} \sqrt{\frac{\lambda_{c,i} + \sigma^2(t)}{\lambda_{c,i} + \sigma^2(T)}} \mathbf{u}_{c,i}^T (\mathbf{x}_T - \mu_c) \mathbf{u}_{c,i} + \gamma \mathbf{U}_c \mathbf{B}_{\sigma(t)} \mathbf{U}_c^T (\mu_c - \mu_{uc}),$$

where  $h(\lambda_{c,i}, \lambda_{uc,i}) = \frac{\lambda_{c,i} + \sigma^2(t)}{\lambda_{c,i} + \sigma^2(T)} \cdot \frac{\lambda_{uc,i} + \sigma^2(T)}{\lambda_{uc,i} + \sigma^2(t)}$  and  $\mathbf{B}_{\sigma_t} = \text{diag}(b_{\sigma(t),1}, \dots, b_{\sigma(t),d})$  has diagonal entries  $b_{\sigma(t),i}$  depending only on  $\lambda_{uc,i}, \lambda_{c,i}$  and  $\sigma(t)$ .

The proof is postponed to appendix B. Compared to the solution of the standard conditional sampling (8), the CFG guidance introduces the following two effects:

- **CPC guidance**  $g_{cpc}(\mathbf{x}_t, t)$  introduces an additional scaling factor  $h(\lambda_{c,i}, \lambda_{uc,i})^{\frac{\gamma}{2}}$  for each component  $\mathbf{u}_{c,i}$  of  $\mathbf{x}_t$ . Since  $h(\lambda_{c,i}, \lambda_{uc,i}) \geq 1$  only if  $\lambda_{c,i} \geq \lambda_{uc,i}$ , the positive CPCs are enhanced. Conversely, the negative CPCs are suppressed. The guidance strength  $\gamma$  serves as an additional control over the degree of enhancement or suppression.
- **Mean-shift guidance term**  $g_{mean}(t)$  shifts  $\mathbf{x}_t$  by  $\gamma \mathbf{U}_c \mathbf{B}_{\sigma(t)} \mathbf{U}_c^T (\mu_c - \mu_{uc})$ , a direction determined by the class-conditional mean offset  $\mu_c - \mu_{uc}$ . Crucially, this shift is independent of the initial noise  $\mathbf{x}_T$  (and intermediate state  $\mathbf{x}_t$ ) and is thus applied consistently to all samples, promoting canonical class features but reducing diversity.

**Empirical Verification.** In Appendix B.3, we provide an empirical validation of Theorem 1 using a 2D synthetic dataset that satisfies Assumption 1. Here, we further verify the CFG’s effects of enhancing (suppressing) CPC components and shifting samples towards the mean-shift direction in natural image dataset through the following experiment:

- For a chosen class, generate 1,000 samples using naive conditional sampling (denoted by  $\mathbf{x}_c$ ) and 1,000 samples using CFG (denoted by  $\mathbf{x}_{cfg}$ ), and center both sets by subtracting the class mean  $\mu_c$ .
- Project each sample onto the positive CPCs (denoted as  $\mathbf{v}_+$ ), the negative CPCs (denoted by  $\mathbf{v}_-$ ), and the mean-shift vector (denoted by  $\mu_c - \mu_{uc}$ ) to obtain a series of univariate distributions.

The above experiments are conducted on both linear and nonlinear (EDM) diffusion models. The resulting univariate distributions are shown in Figure 4(c). Compared with naive conditional sampling,

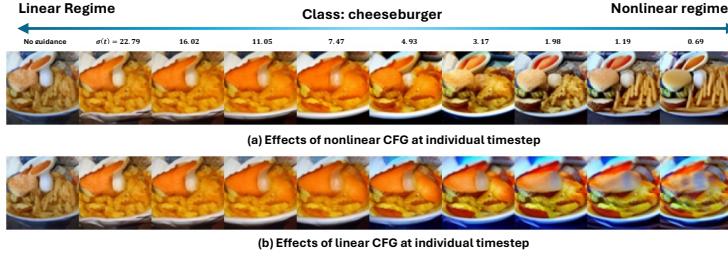


Figure 5: **Linear-to-nonlinear transition in diffusion models.** (a) and (b) compare nonlinear CFG and linear CFG applied to a deep diffusion model (EDM). The leftmost column shows unguided samples; subsequent columns show final samples when guidance is applied only at a specific noise level, with  $\gamma = 15$  (See Figure 25 for more examples).

CFG shifts probability mass toward higher projection values along the positive-CPC and mean-shift directions, and toward lower values along the negative-CPC direction, indicating that the first two are amplified whereas the third is suppressed.

#### 4 Investigating CFG in Nonlinear Models

We now explore how the findings from the linear analysis extend to real-world diffusion models. Recent studies [32, 21, 22] show that diffusion models transition from a linear regime to a nonlinear regime as the noise level decreases. In the linear regime, where  $\sigma(t)$  is large, the learned diffusion denoisers  $\mathcal{D}_\theta$  can be well approximated by the optimal linear denoiser  $\mathcal{D}_L$  (7) (see both qualitative and quantitative verification in appendix F.1). As  $\sigma(t)$  decreases, the diffusion model enters the nonlinear regime where  $\mathcal{D}_\theta$  diverges from  $\mathcal{D}_L$ . Interestingly, this linear-to-nonlinear transition correlates with the coarse-to-fine effects of CFG. As shown in Figures 5 and 25, in the linear regime, linear and nonlinear CFG produce similar effects, substantially reshaping the global structure of the samples. In contrast, in the nonlinear regime, nonlinear CFG primarily refines local details while preserving the overall structure, leading to different effects as those of linear CFG. This linear-to-nonlinear, coarse-to-fine transition motivates our separate analyses of CFG behavior in each regime.

##### 4.1 CFG in the Linear Regime

Figure 6 illustrates the effects of separately applying (i) nonlinear CFG, (ii) linear CFG, (iii) mean-shift guidance, (iv) positive CPC guidance, and (v) negative CPC guidance within the linear regime of EDM over a broad range of  $\gamma$ . As expected, linear CFG<sup>2</sup> produces results that closely match those of nonlinear CFG, both significantly altering the overall structures of unguided samples. Notably, decomposing linear CFG provides further insights:

**Mean-shift guidance dominates CFG in the linear regime.** As shown in Figure 6(a), qualitatively, mean-shift guidance alone replicates the effects of both linear and nonlinear CFG. Consistent with this observation, FD<sub>DINOv2</sub> scores confirm that the mean-shift term is the main contributor to CFG's overall behavior. Because mean-shift term is independent of the sampling trajectory, it can reduce sample diversity. As shown in Figure 6(b), mean-shift guidance improves generation quality only within a limited range of  $\gamma$ , after which further increases in  $\gamma$  degrade FD<sub>DINOv2</sub> scores, reflecting a loss of diversity.

**CPC guidance also improves generation quality.** Although overshadowed by the mean-shift term, applying CPC guidance independently offers notable benefits as well. Qualitatively, posi-

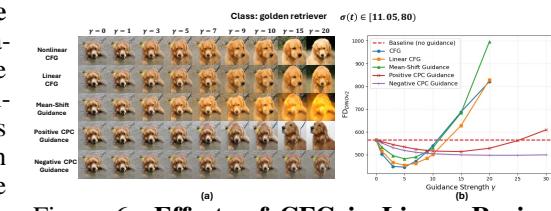


Figure 6: **Effects of CFG in Linear Regime.** Each row in (a) demonstrates the impact of different guidance types applied to EDM within the linear regime (specified in the subtitles), with varying guidance strength  $\gamma$ . (b) shows the FD<sub>DINOv2</sub> scores computed over 50,000 samples. The reported values are relatively high because the scores are computed separately per class, which often has limited number of training images. It is well known that FD<sub>DINOv2</sub> scores can appear inflated when the reference dataset size is small.

<sup>2</sup>Note that the “linear CFG” here differs from the “linear CFG” in section 3, where both the naive conditional score and the cfg guidance are linear. In contrast, the linear guidance in this section, along with its components, is applied to a real-world *deep* diffusion model.

tive and negative CPC terms preserve the global structure of unguided samples while refining existing features, remaining effective over a broader range of  $\gamma$ . Moreover, CPC guidance sometimes mitigate the artifacts introduced by the mean-shift term, such as color oversaturation in the golden retriever example at  $\gamma = 20$ . Lastly, we note that the effects of CPC guidance can vary by class. As shown in Figure 30, negative CPC term improves  $FD_{DINOv2}$  scores for "golden retriever" but has minimal effect on "sports car". These findings are verified on 10 classes, with additional results presented in appendix F.2, Figures 27 to 30.

## 4.2 CFG in the Nonlinear Regime

In the nonlinear regime where  $\sigma(t)$  is small, as shown in Figure 5(b), the effects of linear CFG diverge from those of the actual nonlinear CFG. By Tweedie's formula, the CFG guidance (5) can be expressed as  $g(\mathbf{x}, t) = \frac{\mathcal{D}(\mathbf{x}; \sigma(t), \mathbf{c}) - \mathcal{D}(\mathbf{x}; \sigma(t))}{\sigma^2(t)}$ , where  $\mathcal{D}(\mathbf{x}; \sigma(t), \mathbf{c})$  and  $\mathcal{D}(\mathbf{x}; \sigma(t))$  denote the optimal conditional and unconditional denoisers minimizing (3). Unlike in the linear setting, these denoisers do not admit closed-form expressions in the nonlinear regime, making analytical study difficult. Nevertheless, when denoisers are parameterized by deep networks with no additive 'bias' parameters, their input-output mappings are locally piecewise linear [33, 34], satisfying:

$$\mathcal{D}(\mathbf{x}; \sigma(t), \mathbf{c}) = \nabla_{\mathbf{x}} \mathcal{D}(\mathbf{x}; \sigma(t), \mathbf{c}) \mathbf{x}, \quad \mathcal{D}(\mathbf{x}; \sigma(t)) = \nabla_{\mathbf{x}} \mathcal{D}(\mathbf{x}; \sigma(t)) \mathbf{x}, \quad (16)$$

where  $\nabla_{\mathbf{x}} \mathcal{D}(\mathbf{x}; \sigma(t), \mathbf{c})$  and  $\nabla_{\mathbf{x}} \mathcal{D}(\mathbf{x}; \sigma(t))$  are the local Jacobians of the denoisers. In this case, the CFG guidance becomes  $\frac{(\nabla_{\mathbf{x}} \mathcal{D}(\mathbf{x}; \sigma(t), \mathbf{c}) - \nabla_{\mathbf{x}} \mathcal{D}(\mathbf{x}; \sigma(t))) \mathbf{x}}{\sigma^2(t)}$ , which shares a similar form as  $g_{cpc}(\mathbf{x}, t)$  defined under the linear setting in (12), since  $\tilde{\Sigma}_{c,t} - \tilde{\Sigma}_{uc,t} = \nabla_{\mathbf{x}} \mathcal{D}_L(\mathbf{x}; \sigma(t), \mathbf{c}) - \nabla_{\mathbf{x}} \mathcal{D}_L(\mathbf{x}; \sigma(t))$ , where  $\mathcal{D}_L$  is the optimal linear denoiser. Thus, the guidance can again be decomposed into positive and negative CPC components, enhancing the former and suppressing the latter. The key distinction from the linear setting is that here, the CPCs are adaptive to the current input  $\mathbf{x}$ .

The bias-free denoisers belong to the broader class of pseudo-linear denoisers [35, 36], which admit the form  $\mathcal{D}(\mathbf{x}; \sigma(t)) = \mathbf{W}(\mathbf{x}; \sigma(t)) \mathbf{x}$ , where  $\mathbf{W}(\mathbf{x}; \sigma(t))$  is symmetric and input-dependent. Importantly, it is shown in [35] that if the origin is a stationary point of the log-density, i.e.,  $\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma(t))|_{\mathbf{x}=0} = 0$ , then the optimal denoiser must possess such a piecewise linear structure. Even if the diffusion models are not bias-free and the locally linear property does not hold exactly, (16) still serves as a reasonable proxy. As discussed in section 2.3, the Jacobian  $\nabla_{\mathbf{x}} \mathcal{D}(\mathbf{x}; \sigma(t))$  is proportional to the posterior covariance. Its leading singular vectors capture the dominant structures shared by all plausible clean images corresponding to the noisy input  $\mathbf{x}$ , while directions associated with near-zero singular values span a null space irrelevant to the image content. Hence, (16) performs a weighted projection onto the subspace encoding the most informative image structures—effectively functioning as a valid denoising operator. Indeed, as shown in Figure 7(b)–(c), both conditional and unconditional Jacobians effectively denoise the input, although their outputs appear brighter and sharper than those from the actual denoisers in (d)–(e). Comparing Figure 7(b) and (c), we find that the conditional and unconditional Jacobians yield denoised outputs with similar global structure, which implies both capture the generic structure of the current sample. However, the conditional Jacobian additionally preserves finer, class-specific details. A similar pattern holds for the actual denoisers shown in Figure 7(d) and (e).

For guidance purposes, our goal is to selectively enhance these fine, class-specific details that the conditional denoiser captures but the unconditional one does not. Achieving this requires identifying directions that encode class-dependent information from those represent generic structures. Empirically, as shown in Figure 8, the following guidance, inspired by the positive CPC guidance (13), can lead to similar effects as the actual nonlinear CFG, sharpening image details:

$$\frac{\gamma}{\sigma^2(t)} \sum_i \hat{\lambda}_{+,i} \mathbf{v}_{+,i} (\mathbf{v}_{+,i}^T \nabla \mathcal{D}_{\theta}(\mathbf{x}_t; \sigma(t), \mathbf{c})), \quad (17)$$

where  $\hat{\lambda}_{+,i}$  and  $\mathbf{v}_{+,i}$  denote the positive eigenvalues and eigenvectors of  $\nabla \mathcal{D}_{\theta}(\mathbf{x}_t; \sigma(t), \mathbf{c}) - \nabla \mathcal{D}_{\theta}(\mathbf{x}_t; \sigma(t))$ . Unlike (13), this guidance applies projection to the denoiser's output rather than the

noisy input  $\mathbf{x}_t$ , which we find leads to better qualitative results. For comparison, we also test the following non-selective guidance that enhances all conditional posterior PCs:

$$\frac{\gamma}{\sigma^2(t)} \sum_i \lambda_{c,i} \mathbf{u}_{c,i} (\mathbf{u}_{c,i}^T \mathcal{D}_{\theta}(\mathbf{x}_t; \sigma(t), \mathbf{c})), \quad (18)$$

where  $\lambda_{c,i}$  and  $\mathbf{u}_{c,i}$  are the singular values and vectors of  $\nabla \mathcal{D}_{\theta}(\mathbf{x}_t; \sigma(t), \mathbf{c})$ . As shown in Figure 8, this approach fails to improve image quality and frequently produces images with oversaturated colors, indicating that not all conditional posterior PCs correspond to class-specific features—effective guidance must selectively amplify only those that do.

We note that our heuristic guidance serves as a conceptual approximation and may not always perfectly align with actual CFG behavior; in practice, the actual nonlinear CFG yields more stable and consistent results. Due to the black-box nature of deep networks, fully characterizing this mechanism remains challenging, and we regard this as an important direction for future research.

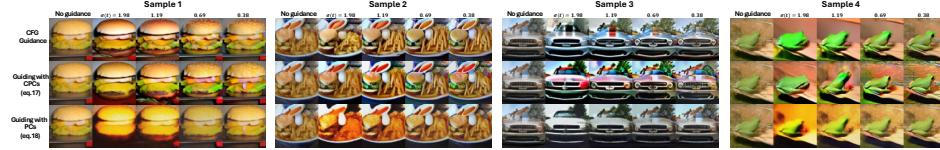


Figure 8: **Effects of CFG in the Nonlinear Regime.** Different guidance methods, each with a fixed strength of  $\gamma = 15$ , are applied at individual timesteps in the nonlinear regime. Each image shows the final output when guidance is applied solely at the timestep indicated at the top. Note that (17) matches the effects of CFG by enhancing finer image details, whereas (18) does not improve generation quality. For additional experimental results, please refer to Figure 36.

## 5 Discussion and Conclusion

The experiments in the main-text are conducted extensively using the EDM-1 model [4], which operates directly in pixel space with  $64 \times 64$  resolution. In appendix G, we present complementary results on the EDM-2 [5] latent diffusion model, which generates images at  $512 \times 512$  resolution.

The key insight of this work is that CFG enhances generation quality by amplifying class-specific features while suppressing generic ones. In the linear setting, this effect emerges from the interplay of three guidance components. Different from previous works which mainly focus on analyzing isotropic Gaussian distributions, our study probes the covariance structures of image data, revealing that salient class-specific features emerge from contrast between class covariances.

Although our analysis is based on linear diffusion model (Gaussian data) assumption, the results remain noteworthy since: (i) CFG significantly enhances the generation quality of linear diffusion models, making the linear setting a meaningful stand-alone testbed for studying CFG and (ii) real-world diffusion models can be well-approximated by their linear counterparts for a wide range of noise levels. We note that the dynamics of linear setting is by itself complex: an interpretable solution to the linear reverse ODE is unattainable unless additional assumptions are imposed on the covariance structures (e.g., the common principal components assumption). A natural next step is to extend the analysis to Gaussian mixtures. We have made some initial attempts in appendix H, showing that CFG guidance in the Gaussian mixture setting can be decomposed in a similar manner as the single Gaussian case.

We believe our findings open several promising directions for future research. First, the observed lack of class-specificity issue implies the current training procedures for diffusion models remain suboptimal. This highlights the need for developing principled training objectives that explicitly encourage the model to learn class-specific patterns. Second, beyond the context of CFG, PCA has been widely utilized for extracting visual features or semantic concepts from diffusion models [37–39]. Our results suggest that extending these approaches with Contrastive PCA can be a promising next step for more controllable and interpretable generation.

## Data Availability Statement

The code and instructions for reproducing our experiments will be released at: <https://github.com/Morefre/Towards-Understanding-the-Mechanisms-of-Classifier-Free-Guidance.git>.

## Acknowledgment

We acknowledge funding support from NSF CCF-2212066, NSF CCF- 2212326, NSF IIS 2402950, and ONR N000142512339. This research used the Delta advanced computing and data resource which is supported by the National Science Foundation (award OAC 2005572) and the State of Illinois. Delta is a joint effort of the University of Illinois Urbana-Champaign and its National Center for Supercomputing Applications[40]. We thank Mr. Yixiang Dai for fruitful discussions and valuable feedbacks.

## References

- [1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [2] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.
- [3] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.
- [4] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.
- [5] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24174–24184, 2024.
- [6] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.
- [7] Arwen Bradley and Preetum Nakkiran. Classifier-free guidance is a predictor-corrector. *arXiv preprint arXiv:2408.09000*, 2024.
- [8] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.
- [9] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7462–7471, 2023.
- [10] Tuomas Kynkänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *arXiv preprint arXiv:2404.07724*, 2024.
- [11] Tero Karras, Miika Aittala, Tuomas Kynkänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *arXiv preprint arXiv:2406.02507*, 2024.
- [12] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. In *The Twelfth International Conference on Learning Representations*, 2024.

[13] Seyedmorteza Sadat, Manuel Kansy, Otmar Hilliges, and Romann M Weber. No training, no problem: Rethinking classifier-free guidance for diffusion models. *arXiv preprint arXiv:2407.02687*, 2024.

[14] Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*, 2024.

[15] Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, pages 1–17. Springer, 2025.

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[17] 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.

[18] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[19] Yuchen Wu, Minshuo Chen, Zihao Li, Mengdi Wang, and Yuting Wei. Theoretical insights for diffusion guidance: A case study for gaussian mixture models. In *Forty-first International Conference on Machine Learning*, 2024.

[20] Muthu Chidambaram, Khashayar Gatmiry, Sitan Chen, Holden Lee, and Jianfeng Lu. What does guidance do? a fine-grained analysis in a simple setting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[21] Xiang Li, Yixiang Dai, and Qing Qu. Understanding generalizability of diffusion models requires rethinking the hidden gaussian structure. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[22] Binxu Wang and John Vastola. The unreasonable effectiveness of gaussian score approximation for diffusion models and its applications. *Transactions on Machine Learning Research*, 2024.

[23] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[24] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[25] Abubakar Abid, Martin J Zhang, Vivek K Bagaria, and James Zou. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature communications*, 9(1):2134, 2018.

[26] Hila Manor and Tomer Michaeli. On the posterior distribution in denoising: Application to uncertainty quantification. In *The Twelfth International Conference on Learning Representations*, 2024.

[27] 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. *Advances in neural information processing systems*, 30, 2017.

[28] Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. *Advances in neural information processing systems*, 32, 2019.

[29] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

- [30] Bernhard N Flury. Common principal components in k groups. *Journal of the American Statistical Association*, 79(388):892–898, 1984.
- [31] Bernhard Flury. *Common principal components & related multivariate models*. John Wiley & Sons, Inc., USA, 1988.
- [32] Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [33] Zahra Kadkhodaie, Florentin Guth, Eero P Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth International Conference on Learning Representations*, 2024.
- [34] Sreyas Mohan, Zahra Kadkhodaie, Eero P Simoncelli, and Carlos Fernandez-Granda. Robust and interpretable blind image denoising via bias-free convolutional neural networks. In *8th International Conference on Learning Representations, ICLR 2020*, 2020.
- [35] Peyman Milanfar and Mauricio Delbracio. Denoising: A powerful building-block for imaging, inverse problems, and machine learning. *arXiv preprint arXiv:2409.06219*, 2024.
- [36] Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (red). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017.
- [37] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36:24129–24142, 2023.
- [38] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475, 2024.
- [39] Rohit Gandikota, Zongze Wu, Richard Zhang, David Bau, Eli Shechtman, and Nick Kolkin. Sliderspace: Decomposing the visual capabilities of diffusion models. *arXiv preprint arXiv:2502.01639*, 2025.
- [40] Timothy J Boerner, Stephen Deems, Thomas R Furlani, Shelley L Knuth, and John Towns. Access: Advancing innovation: Nsf’s advanced cyberinfrastructure coordination ecosystem: Services & support. In *Practice and Experience in Advanced Research Computing 2023: Computing for the Common Good*, pages 173–176. 2023.
- [41] Didong Li, Andrew Jones, and Barbara Engelhardt. Probabilistic contrastive principal component analysis. *arXiv preprint arXiv:2012.07977*, 2020.
- [42] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

# Appendices

---

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Preliminaries</b>	<b>3</b>
2.1	Optimal Linear Diffusion Model	3
2.2	Contrastive Principal Component Analysis	3
2.3	Posterior Data Covariance	4
<b>3</b>	<b>Analyzing CFG in Linear Model</b>	<b>4</b>
3.1	Naive Conditional Generation Lacks Class-Specificity	4
3.2	How Linear CFG Leads to Distinct Generations	5
<b>4</b>	<b>Investigating CFG in Nonlinear Models</b>	<b>8</b>
4.1	CFG in the Linear Regime	8
4.2	CFG in the Nonlinear Regime	9
<b>5</b>	<b>Discussion and Conclusion</b>	<b>10</b>
<b>A</b>	<b>Contrastive Principal Component Analysis</b>	<b>15</b>
<b>B</b>	<b>Analytical Solution to the Reverse Diffusion ODE</b>	<b>16</b>
B.1	Naive Diffusion Reverse ODE	16
B.2	CFG-Guided ODE	17
B.3	Empirical Verification on Synthetic Data.	19
<b>C</b>	<b>Constructing Linear Denoisers</b>	<b>19</b>
<b>D</b>	<b>Naive Conditional Generation Lacks Class-Specificity</b>	<b>20</b>
D.1	Qualitative Results	20
D.2	Quantitative Results	20
D.3	Covariance Matrices of Different Classes Lack Class-Specificity	25
<b>E</b>	<b>Mechanism of Linear CFG</b>	<b>25</b>
E.1	Mean-Shift Guidance	25
E.2	CPC guidance	26
E.3	Distinct Effects of the CFG Components	27
<b>F</b>	<b>CFG in Nonlinear Deep Diffusion Models</b>	<b>34</b>
F.1	Linear to Nonlinear Transition in Diffusion Models	34
F.2	CFG in the Linear Regime	34

F.3	Mean-Shifted Noise Initialization . . . . .	36
F.4	CFG in the Nonlinear Regime . . . . .	36
<b>G</b>	<b>Experimental Results on Latent Diffusion Models</b>	<b>46</b>
<b>H</b>	<b>CFG in Gaussian Mixture Model</b>	<b>46</b>
<b>I</b>	<b>Computing Resources</b>	<b>48</b>

## A Contrastive Principal Component Analysis

Principal component analysis (PCA) finds the features that explain the most variances in the dataset, however, features with high variance do not necessarily correspond to distinct patterns of the target class.

To discover low-dimensional structure that is unique to a dataset, the work [25] proposes the contrastive principal component analysis (CPCA), which includes a *background* (or reference) dataset to highlight patterns unique to the *target* dataset. Let  $X$  and  $Y$  be two datasets with covariance matrices  $\Sigma_X$  and  $\Sigma_Y$ , respectively. For a unit vector  $\mathbf{v} \in \mathbb{S}^{d-1}$ , its variances  $\text{Var}_X(\mathbf{v})$  and  $\text{Var}_Y(\mathbf{v})$  in the two datasets are:

$$\text{Var}_X(\mathbf{v}) := \mathbf{v}^T \Sigma_X \mathbf{v}, \quad \text{Var}_Y(\mathbf{v}) := \mathbf{v}^T \Sigma_Y \mathbf{v}. \quad (19)$$

If  $\mathbf{v}$  corresponds to a unique class-specific pattern of  $X$ , we expect  $\text{Var}_X(\mathbf{v}) \gg \text{Var}_Y(\mathbf{v})$ , i.e. it explains significantly more variance in  $X$  than in  $Y$ . Such directions, called the *contrastive principal components* (CPCs), can be found by iteratively solving:

$$\arg \max_{\mathbf{v} \in \mathbb{S}^{d-1}} \mathbf{v}^T (\Sigma_X - \Sigma_Y) \mathbf{v}, \quad (20)$$

where, at each iteration, the resulting  $\mathbf{v}$  is subtracted from  $\Sigma_X - \Sigma_Y$ . These directions are essentially the eigenvectors of  $\Sigma(X) - \Sigma(Y)$ . Conversely, directions  $\mathbf{v}$  for which  $\text{Var}_X(\mathbf{v}) \approx \text{Var}_Y(\mathbf{v})$  represent either universal structures shared by both  $X$  and  $Y$  or meaningless features lying in the null space of the data covariances—and are thus discarded as less interesting.

**Geometric Interpretation of CPCA.** Geometrically, the first  $k$  CPCs span the  $k$ -dimensional subspace that best fits the dataset  $X$  while being as far as possible from  $Y$  [41]. This is proved by the following theorem:

**Theorem 2.** *Without loss of generality, assume  $p_X(\mathbf{x})$  and  $p_Y(\mathbf{y})$  have zero means (i.e., the data is centered). Then the following objective is equivalent to (20):*

$$\arg \min_{\mathbf{v} \in \mathbb{S}^{d-1}} \mathbb{E}_{\mathbf{x} \sim p_X} \|\mathbf{x} - \mathbf{v}\mathbf{v}^T \mathbf{x}\|_2^2 - \mathbb{E}_{\mathbf{y} \sim p_Y} \|\mathbf{y} - \mathbf{v}\mathbf{v}^T \mathbf{y}\|_2^2. \quad (21)$$

*Proof:*

$$\begin{aligned} \mathbf{v} &= \arg \min_{\mathbf{v} \in \mathbb{S}^{d-1}} \mathbb{E}_{\mathbf{x} \sim p_X} \|\mathbf{x} - \mathbf{v}\mathbf{v}^T \mathbf{x}\|_2^2 - \mathbb{E}_{\mathbf{y} \sim p_Y} \|\mathbf{y} - \mathbf{v}\mathbf{v}^T \mathbf{y}\|_2^2 \\ &= \arg \min_{\mathbf{v} \in \mathbb{S}^{d-1}} \mathbb{E}_{\mathbf{x}} (\mathbf{x}^T \mathbf{x} - \mathbf{x}^T \mathbf{v} \mathbf{v}^T \mathbf{x}) - \mathbb{E}_{\mathbf{y}} (\mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{v} \mathbf{v}^T \mathbf{y}) \\ &= \arg \min_{\mathbf{v} \in \mathbb{S}^{d-1}} \mathbb{E}_{\mathbf{x}} (-\mathbf{x}^T \mathbf{v} \mathbf{v}^T \mathbf{x}) - \mathbb{E}_{\mathbf{y}} (-\mathbf{y}^T \mathbf{v} \mathbf{v}^T \mathbf{y}) \\ &= \arg \min_{\mathbf{v} \in \mathbb{S}^{d-1}} -\mathbb{E}_{\mathbf{x}} (\mathbf{x}^T \mathbf{v} \mathbf{v}^T \mathbf{x}) + \mathbb{E}_{\mathbf{y}} (\mathbf{y}^T \mathbf{v} \mathbf{v}^T \mathbf{y}) \\ &= \arg \max_{\mathbf{v} \in \mathbb{S}^{d-1}} \mathbb{E}_{\mathbf{x}} (\mathbf{x}^T \mathbf{v} \mathbf{v}^T \mathbf{x}) - \mathbb{E}_{\mathbf{y}} (\mathbf{y}^T \mathbf{v} \mathbf{v}^T \mathbf{y}) \\ &= \arg \max_{\mathbf{v} \in \mathbb{S}^{d-1}} \mathbf{v}^T \mathbb{E}_{\mathbf{x}} (\mathbf{x} \mathbf{x}^T) \mathbf{v} - \mathbf{v}^T \mathbb{E}_{\mathbf{y}} (\mathbf{y} \mathbf{y}^T) \mathbf{v} \\ &= \arg \max_{\mathbf{v} \in \mathbb{S}^{d-1}} \mathbf{v}^T (\Sigma_X - \Sigma_Y) \mathbf{v}. \end{aligned}$$

This proof is adapted from [41].

## B Analytical Solution to the Reverse Diffusion ODE

In this section, we examine the solutions to both the naive reverse diffusion ODE (2) and the CFG-guided reverse diffusion ODE (6) in the context of linear diffusion models. We follow the EDM formulation [4], which uses time schedule  $\sigma(t) = t$ . Notice that this same schedule is also used by the well-known DDIM sampler [2].

### B.1 Naive Diffusion Reverse ODE

We begin by analyzing the diffusion ODE (2) with no guidance applied. The proof below is borrowed from [22].

Let  $\mu$  and  $\Sigma$  be the mean and covariance of  $p_{\text{data}}(\mathbf{x})$  respectively. Suppose  $\mathbf{U}\Lambda\mathbf{U}^T$  is the full SVD of  $\Sigma$  with  $\mathbf{U} \in \mathbb{R}^{d \times d}$  being orthonormal and  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$  contains the singular values. For image datasets, the covariance is often low-rank implying some singular values are 0. Under the constraint that  $\mathcal{D}(\mathbf{x}; \sigma(t))$  is linear (with a bias term), the optimal solution to (3) has the closed-form [21]:

$$\mathcal{D}(\mathbf{x}; \sigma(t)) = \mu + \mathbf{U}\tilde{\Lambda}_{\sigma(t)}\mathbf{U}^T(\mathbf{x} - \mu), \quad (22)$$

where  $\tilde{\Lambda}_{\sigma(t)} = \text{diag}\left(\frac{\lambda_1}{\lambda_1 + \sigma^2(t)}, \dots, \frac{\lambda_d}{\lambda_d + \sigma^2(t)}\right)$ . This optimal linear solution is obtained by setting the derivative of (3) with respect to the weight and bias to zero, leveraging the fact that the objective is convex under the linear constraint.

Following the EDM framework, this optimal linear denoiser yields the sampling trajectory for (2) as:

$$d\mathbf{x} = -\sigma \nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma) d\sigma \quad (23)$$

$$\Rightarrow d\mathbf{x} = \frac{(\mathbf{I} - \mathbf{U}\tilde{\Lambda}_{\sigma}\mathbf{U}^T)(\mathbf{x} - \mu)}{\sigma} d\sigma \quad (24)$$

$$\Rightarrow d(\mathbf{x} - \mu) = \frac{\mathbf{U}(\mathbf{I} - \tilde{\Lambda}_{\sigma})\mathbf{U}^T(\mathbf{x} - \mu)}{\sigma} d\sigma, \quad (25)$$

where we omit the subscript  $t$  for simplicity.

Define  $c_k(\sigma) = \mathbf{u}_k^T(\mathbf{x} - \mu)$  for  $k \in \{1, \dots, d\}$ , we have:

$$dc_k(\sigma) = \frac{\sigma}{\lambda_k + \sigma^2} c_k(\sigma) d\sigma \quad (26)$$

$$\Rightarrow \frac{dc_k(\sigma)}{c_k(\sigma)} = \frac{\sigma}{\lambda_k + \sigma^2} d\sigma. \quad (27)$$

Integrating both sides of (27), we get:

$$d \log c_k(\sigma) = d \log \sqrt{\lambda_k + \sigma^2} \quad (28)$$

$$\Rightarrow c_k(\sigma) = \sqrt{\lambda_k + \sigma^2} C, \quad (29)$$

where  $C$  is the integral constant. Using the initial condition  $c_k(\sigma_T) = \mathbf{u}_k^T(\mathbf{x}_T - \mu)$ , we have:

$$C = \frac{\mathbf{u}_k^T(\mathbf{x}_T - \mu)}{\sqrt{\lambda_k + \sigma_T^2}} \quad (30)$$

$$\Rightarrow c_k(\sigma) = \sqrt{\frac{\lambda_k + \sigma^2}{\lambda_k + \sigma_T^2}} \mathbf{u}_k^T(\mathbf{x}_T - \mu) \quad (31)$$

$$\Rightarrow \mathbf{x}_t = \mu + \sum_{k=1}^d \sqrt{\frac{\lambda_k + \sigma_t^2}{\lambda_k + \sigma_T^2}} \mathbf{u}_k^T(\mathbf{x}_T - \mu) \mathbf{u}_k, \quad (32)$$

where the last equality holds because  $\mathbf{x}_t - \mu = \sum_{i=k}^d c_k(\sigma_t) \mathbf{u}_k$

Notice that the generated samples are primarily determined by the data's covariance structure. However, since the covariance may not capture the most distinctive features of a specific class, the resulting images often lack sufficient class specificity.

## B.2 CFG-Guided ODE

To apply CFG, we need two separate models corresponding to conditional and unconditional data respectively. Let  $\mu_c$ ,  $\mu_{uc}$  be the means of the conditional and unconditional data, and let  $\Sigma_c = \mathbf{U}_c \Lambda_c \mathbf{U}_c^T$ ,  $\Sigma_{uc} = \mathbf{U}_{uc} \Lambda_{uc} \mathbf{U}_{uc}^T$  be their corresponding covariances, where  $\Lambda_c = \text{diag}(\lambda_{c,1}, \dots, \lambda_{c,d})$  and  $\Lambda_{uc} = \text{diag}(\lambda_{uc,1}, \dots, \lambda_{uc,d})$ . Then the conditional and unconditional optimal linear denoisers take the following forms:

$$\mathcal{D}_L(\mathbf{x}; \sigma(t), \mathbf{c}) = \mu_c + \mathbf{U}_c \tilde{\Lambda}_{c,\sigma(t)} \mathbf{U}_c^T (\mathbf{x} - \mu_c), \quad (33)$$

$$\mathcal{D}_L(\mathbf{x}; \sigma(t)) = \mu_{uc} + \mathbf{U}_{uc} \tilde{\Lambda}_{uc,\sigma(t)} \mathbf{U}_{uc}^T (\mathbf{x} - \mu_{uc}), \quad (34)$$

Then the CFG sampling trajectory (6) can be expressed in terms of the optimal linear denoisers:

$$d\mathbf{x}_t = -\sigma(t) \left( \frac{\mathcal{D}_L(\mathbf{x}_t; \sigma(t), \mathbf{c}) - \mathbf{x}_t}{\sigma^2(t)} + \gamma \frac{\mathcal{D}_L(\mathbf{x}_t; \sigma(t), \mathbf{c}) - \mathcal{D}_L(\mathbf{x}_t; \sigma(t))}{\sigma^2(t)} \right) dt \quad (35)$$

$$= -\frac{1}{\sigma(t)} (\mathbf{U}_c \tilde{\Lambda}_{\sigma(t),c} \mathbf{U}_c^T - \mathbf{I}) (\mathbf{x}_t - \mu_c) dt \quad (36)$$

$$- \frac{\gamma}{\sigma(t)} (\mathbf{U}_c \tilde{\Lambda}_{\sigma(t),c} \mathbf{U}_c^T - \mathbf{U}_{uc} \tilde{\Lambda}_{\sigma(t),uc} \mathbf{U}_{uc}^T) (\mathbf{x}_t - \mu_c) dt \quad (37)$$

$$- \frac{\gamma}{\sigma(t)} (\mathbf{I} - \mathbf{U}_{uc} \tilde{\Lambda}_{\sigma(t),uc} \mathbf{U}_{uc}^T) (\mu_c - \mu_{uc}) dt, \quad (38)$$

where (36) is the naive conditional score while (37) and (38) together form the CFG guidance direction. Note that under the setting of linear diffusion model, we have

$$p(\mathbf{x}; \sigma(t)) = \mathcal{N}(\mu_{uc}, \Sigma_{uc} + \sigma^2(t)\mathbf{I}), \quad (39)$$

$$p(\mathbf{x}|\mathbf{c}; \sigma(t)) = \mathcal{N}(\mu_c, \Sigma_c + \sigma^2(t)\mathbf{I}), \quad (40)$$

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma(t)) = (\Sigma_{uc} + \sigma^2(t)\mathbf{I})^{-1} (\mu_{uc} - \mathbf{x}), \quad (41)$$

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{c}; \sigma(t)) = (\Sigma_c + \sigma^2(t)\mathbf{I})^{-1} (\mu_c - \mathbf{x}), \quad (42)$$

$$\mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{c}; \sigma(t))} [\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{c}; \sigma(t)) - \nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma(t))] = (\Sigma_{uc} + \sigma^2(t)\mathbf{I})^{-1} (\mu_c - \mu_{uc}) \quad (43)$$

$$= \frac{1}{\sigma^2(t)} (\mathbf{I} - \tilde{\Sigma}_{uc,t}) (\mu_c - \mu_{uc}). \quad (44)$$

Therefore, we have:

$$g_{mean}(t) = \gamma \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{c}; \sigma(t))} [\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{c}; \sigma(t)) - \nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma(t))] \quad (45)$$

$$= \gamma \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|\mathbf{c}; \sigma(t))} [\nabla_{\mathbf{x}} \log \frac{p(\mathbf{x}|\mathbf{c}; \sigma(t))}{p(\mathbf{x}; \sigma(t))}] \quad (46)$$

which implies the mean-shift guidance term can be interpreted as the probability-weighted average of the steepest ascent direction that maximizes the log-likelihood ratio of the noise mollified conditional and unconditional distributions.

Since (35) is a first-order non-homogeneous differential equation, its closed-form solution can in principle, be expressed through integrals. However, these integrals cannot be explicitly evaluated or decoupled in the general case. To obtain a tractable, interpretable solution, we must impose additional assumptions on the structures of  $\Sigma_c$  and  $\Sigma_{uc}$ . Therefore, we make the following assumptions:

**Assumption.** *The covariance matrices  $\Sigma_c$  and  $\Sigma_{uc}$  are simultaneously diagonalizable, i.e.,  $\Sigma_{uc} = \mathbf{U}_c \Lambda_{uc} \mathbf{U}_c^T$ , where  $\mathbf{U}_c \in \mathbb{R}^{d \times d}$  are the principal components (singular vectors) of the conditional data. Here  $\Lambda_{uc}$  is not necessarily ordered by the magnitude of the singular value.*

This is well-known as the *Common Principal Components assumption* [30, 31], widely utilized to analyze structural relationships across data groups. Under this assumption, the relative importance of the  $k^{\text{th}}$  principal component  $\mathbf{u}_{c,k}$  in the conditional and unconditional datasets is fully determined by the relative magnitudes of its associated singular values:

- If  $\lambda_{c,k} > \lambda_{uc,k}$ , then  $\mathbf{u}_{c,k}$  explains more variance in the conditional dataset than in the unconditional dataset, i.e., it is more distinct for the conditional data. This corresponds to the positive CPC discussed in the main text.

- Conversely, if  $\lambda_{c,k} < \lambda_{uc,k}$ , then  $\mathbf{u}_{c,k}$  explains more variance in the unconditional dataset than in the conditional dataset, making it more distinct for the unconditional data and therefore it is a negative CPC.

Under the assumption, the CFG guided ODE (35) can be simplified as:

$$d\mathbf{x} = \frac{(\mathbf{I} - \mathbf{U}_c \tilde{\Lambda}_{\sigma,c} \mathbf{U}_c^T)(\mathbf{x} - \boldsymbol{\mu}_c)}{\sigma} d\sigma - \gamma \frac{\mathbf{U}_c (\tilde{\Lambda}_{\sigma,c} - \tilde{\Lambda}_{\sigma,uc}) \mathbf{U}_c^T (\mathbf{x} - \boldsymbol{\mu}_c)}{\sigma} d\sigma \quad (47)$$

$$- \gamma \frac{\mathbf{U}_c (\mathbf{I} - \tilde{\Lambda}_{\sigma,uc}) \mathbf{U}_c^T (\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc})}{\sigma} d\sigma. \quad (48)$$

Define  $c_k(\sigma) = \mathbf{u}_{c,k}^T (\mathbf{x} - \boldsymbol{\mu}_c)$ , we have:

$$dc_k(\sigma) = \frac{\sigma}{\lambda_{c,k} + \sigma^2} c_k(\sigma) d\sigma - \gamma \frac{\sigma(\lambda_{c,k} - \lambda_{uc,k})}{(\lambda_{c,k} + \sigma^2)(\lambda_{uc,k} + \sigma^2)} c_k(\sigma) d\sigma \quad (49)$$

$$- \gamma \frac{\sigma}{\lambda_{uc,k} + \sigma^2} \mathbf{u}_{c,k}^T (\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc}) d\sigma. \quad (50)$$

Therefore, the dynamics of  $dc_k(\sigma)$  can be expressed as:

$$dc_k(\sigma) + f(\sigma) c_k(\sigma) d\sigma = g(\sigma) d\sigma, \quad (51)$$

, where  $f(\sigma) = -(\frac{\sigma}{\lambda_{c,k} + \sigma^2} - \gamma \frac{\sigma(\lambda_{c,k} - \lambda_{uc,k})}{(\lambda_{c,k} + \sigma^2)(\lambda_{uc,k} + \sigma^2)})$  and  $g(\sigma) = -\gamma \frac{\sigma}{\lambda_{uc,k} + \sigma^2} \mathbf{u}_{c,k}^T (\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc})$ .

**Homogeneous ODE.** We first consider the homogeneous counterpart of (51):

$$dc_k(\sigma) = -f(\sigma) c_k(\sigma) d\sigma, \quad (52)$$

where  $-f(\sigma) c_k(\sigma) d\sigma$  corresponds to the combination of the standard conditional score (36) and the CPC guidance term (37). Integrating over both sides of (52), we get:

$$c_k(\sigma) = C e^{\int -f(\sigma) d\sigma}. \quad (53)$$

Notice that:

$$\int -f(\sigma) d\sigma = \int \frac{\sigma}{\lambda_{c,k} + \sigma^2} d\sigma - \gamma \int \frac{\sigma(\lambda_{c,k} - \lambda_{uc,k})}{(\lambda_{c,k} + \sigma^2)(\lambda_{uc,k} + \sigma^2)} d\sigma \quad (54)$$

$$= \frac{1}{2} \ln(\lambda_{c,k} + \sigma^2) + \frac{\gamma}{2} \ln\left(\frac{\lambda_{c,k} + \sigma^2}{\lambda_{uc,k} + \sigma^2}\right), \quad (55)$$

which implies:

$$c_k(\sigma) = C (\lambda_{c,k} + \sigma^2)^{\frac{1}{2}} \left(\frac{\lambda_{c,k} + \sigma^2}{\lambda_{uc,k} + \sigma^2}\right)^{\frac{\gamma}{2}}. \quad (56)$$

Applying the initial condition that  $c_k(\sigma_T) = \mathbf{u}_{c,k}^T (\mathbf{x}_T - \boldsymbol{\mu}_c)$ , we have:

$$C = (\lambda_{c,k} + \sigma_T^2)^{-\frac{1}{2}} \left(\frac{\lambda_{c,k} + \sigma_T^2}{\lambda_{uc,k} + \sigma_T^2}\right)^{-\frac{\gamma}{2}} \mathbf{u}_{c,k}^T (\mathbf{x}_T - \boldsymbol{\mu}_c) \quad (57)$$

(58)

$$\Rightarrow c_k(\sigma_t) = \sum_{k=1}^d \left(\frac{\lambda_{c,k} + \sigma_t^2}{\lambda_{c,k} + \sigma_T^2} \frac{\lambda_{uc,k} + \sigma_T^2}{\lambda_{uc,k} + \sigma_t^2}\right)^{\frac{\gamma}{2}} \sqrt{\frac{\lambda_{c,k} + \sigma_t^2}{\lambda_{c,k} + \sigma_T^2}} \mathbf{u}_{c,k}^T (\mathbf{x}_T - \boldsymbol{\mu}_c) \quad (59)$$

$$\Rightarrow \mathbf{x}_t = \boldsymbol{\mu}_c + \sum_{k=1}^d h(\lambda_{c,k}, \lambda_{uc,k})^{\frac{\gamma}{2}} \sqrt{\frac{\lambda_{c,k} + \sigma_t^2}{\lambda_{c,k} + \sigma_T^2}} \mathbf{u}_{c,k}^T (\mathbf{x}_T - \boldsymbol{\mu}_c) \mathbf{u}_{c,k}, \quad (60)$$

where  $h(\lambda_{c,k}, \lambda_{uc,k}) = \frac{\lambda_{c,k} + \sigma(t)^2}{\lambda_{c,k} + \sigma^2(t)} \frac{\lambda_{uc,k} + \sigma^2(t)}{\lambda_{uc,k} + \sigma^2(t)}$ . Compared with the solution to the naive reverse process with no guidance (32), each component of  $\mathbf{x}_t$  differs only by a scalar factor  $h(\lambda_{c,k}, \lambda_{uc,k})^{\frac{\gamma}{2}}$ . Specifically:

- $h(\lambda_{c,k}, \lambda_{uc,k}) \geq 1$  if and only if  $\lambda_{c,k} \geq \lambda_{uc,k}$ , meaning positive CPCs are enhanced (scaled up).
- $h(\lambda_{c,k}, \lambda_{uc,k}) \leq 1$  if and only if  $\lambda_{c,k} \leq \lambda_{uc,k}$ , meaning negative CPCs are suppressed (scaled down).

Note that the guidance strength  $\gamma$  provides additional control, amplifying or reducing the degree of enhancement or suppression for each component.

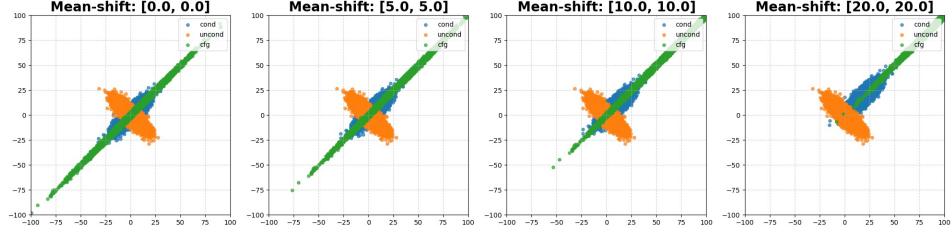


Figure 9: **CFG effects in 2D.** Each subplot differs by the class mean  $\mu_c$ , indicated in the titles. Blue, orange and green points show 1,000 samples generated from conditional sampling, naive unconditional sampling and CFG sampling, respectively.

**Non-Homogeneous ODE.** Let  $\hat{c}_k(\sigma)$  be the solution to the homogeneous ODE (52), then the solution to the non-homogeneous ODE (51) takes the form:

$$c_k(\sigma) = \hat{c}_k(\sigma) + \frac{1}{I(\sigma)} \int I(\sigma') g(\sigma') d\sigma', \quad (61)$$

where  $I(\sigma) = e^{\int f(\sigma') d\sigma'}$  is the integrating factor. Since:

$$I(\sigma) = C(\lambda_{c,k} + \sigma^2)^{-\frac{1}{2}} \left( \frac{\lambda_{c,k} + \sigma^2}{\lambda_{uc,k} + \sigma^2} \right)^{-\frac{\gamma}{2}}, \quad (62)$$

we have:

$$c_k(\sigma) = \hat{c}_k(\sigma) + \gamma(\lambda_{c,k} + \sigma^2)^{\frac{1}{2}} \left( \frac{\lambda_{c,k} + \sigma^2}{\lambda_{uc,k} + \sigma^2} \right)^{\frac{\gamma}{2}} \int_{\sigma}^{\sigma_T} \frac{(\lambda_{uc,k} + \tilde{\sigma}^2)^{\frac{\gamma}{2}-1}}{(\lambda_{c,k} + \tilde{\sigma}^2)^{\frac{\gamma+1}{2}}} \mathbf{u}_{c,k}^T (\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc}) \tilde{\sigma} d\tilde{\sigma} \quad (63)$$

$$= \hat{c}_k(\sigma) + \gamma b_{\sigma,k} \mathbf{u}_{c,k}^T (\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc}), \quad (64)$$

where  $b_{\sigma,k} = (\lambda_{c,k} + \sigma^2)^{\frac{1}{2}} \left( \frac{\lambda_{c,k} + \sigma^2}{\lambda_{uc,k} + \sigma^2} \right)^{\frac{\gamma}{2}} \int_{\sigma}^{\sigma_T} \frac{(\lambda_{uc,k} + \tilde{\sigma}^2)^{\frac{\gamma}{2}-1}}{(\lambda_{c,k} + \tilde{\sigma}^2)^{\frac{\gamma+1}{2}}} \tilde{\sigma} d\tilde{\sigma}$ . Therefore we have:

$$\mathbf{x}_t = \boldsymbol{\mu}_c + \sum_{k=1}^d h(\lambda_{c,k}, \lambda_{uc,k})^{\frac{\gamma}{2}} \sqrt{\frac{\lambda_{c,k} + \sigma_T^2}{\lambda_{c,k} + \sigma_T^2}} \mathbf{u}_{c,k}^T (\mathbf{x}_T - \boldsymbol{\mu}_c) \mathbf{u}_{c,k} + \gamma \sum_{k=1}^d b_k \mathbf{u}_{c,k} \mathbf{u}_{c,k}^T (\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc}) \quad (65)$$

$$= \boldsymbol{\mu}_c + \sum_{k=1}^d h(\lambda_{c,k}, \lambda_{uc,k})^{\frac{\gamma}{2}} \sqrt{\frac{\lambda_{c,k} + \sigma_T^2}{\lambda_{c,k} + \sigma_T^2}} \mathbf{u}_{c,k}^T (\mathbf{x}_T - \boldsymbol{\mu}_c) \mathbf{u}_{c,k} + \gamma \mathbf{U}_c \mathbf{B}_{\sigma_T} \mathbf{U}_c^T (\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc}), \quad (66)$$

where  $\mathbf{B}_{\sigma_T} = \text{diag}(b_{\sigma_T,1}, \dots, b_{\sigma_T,d})$ . Here  $b_k$  depends only on  $\lambda_{uc,k}, \lambda_{c,k}$  and  $\sigma(t)$ . Hence the mean-shift guidance term (38) has the effect of adding constant perturbations that are independent of the initial noise  $\mathbf{x}_T$  to the sampling trajectory.

### B.3 Empirical Verification on Synthetic Data.

We validate Theorem 1 on a 2D toy model with  $\mathbf{U}_c = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}$ ,  $\boldsymbol{\Lambda}_c = \begin{bmatrix} 10 & 0 \\ 0 & 3 \end{bmatrix}$  and  $\boldsymbol{\Lambda}_{uc} = \begin{bmatrix} 3 & 0 \\ 0 & 10 \end{bmatrix}$ . Figure 9 shows the effects of CFG under different class mean  $\mu_c$  (with  $\gamma = 1$  and  $\boldsymbol{\mu}_{uc} = \mathbf{0}$ ). As predicted, CFG enhances variation along the positive CPC  $\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^T$ , suppresses variation along the negative CPC  $\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}^T$ , and shifts samples roughly toward  $\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc}$  at a rate proportional to  $\gamma \|\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc}\|_2$ .

## C Constructing Linear Denoisers

Constructing the linear denoisers (7) requires estimating the data means and covariances. We perform our experiments on CIFAR-10 [42] and ImageNet dataset [29], estimating the linear denoisers for each dataset in different ways:

- **CIFAR-10.** This dataset consists of 10 different classes, each with 5000 images. We obtain the unconditional linear diffusion model by computing the empirical mean and covariance across all 50000 images. For conditional linear diffusion model, we construct a separate linear model for each class, using that class’s mean and covariance estimated from the 5000 images.
- **ImageNet.** This dataset contains 1000 classes, each with approximately 1000 images—a smaller per-class sample size that can introduce bias when estimating means and covariances directly from the training set. Although such direct estimation still yields linear denoisers aligned with the actual diffusion models in the linear regime, these denoisers tend to generate noisier images. We hypothesize that, in the conditional setting, each class’s diffusion denoiser may implicitly leverage information from other classes, meaning the true mean and covariance learned by the deep diffusion model can differ (albeit slightly) from the those estimated solely from that class’s data. To obtain a more accurate linear approximation, we therefore generate 50,000 samples per class with the trained diffusion model, then compute the empirical mean and covariance from these generated samples. Nonetheless, all of our main conclusions remain valid even if we build the linear models using the actual ImageNet training data.

## D Naive Conditional Generation Lacks Class-Specificity

In section 3.1 we argue that naive conditional generation lacks class-specificity and in the linear model setting, such issue can be partially attributed to the non-distinctiveness of the class covariance matrices. In this section, we provide comprehensive experiments to support our claim both qualitatively and quantitatively.

### D.1 Qualitative Results

We generate samples using naive conditional sampling (4) and CFG sampling (6) for all 10 classes of CIFAR-10, as well as for 10 selected ImageNet classes: including (i) class 0: tench , (ii) class 31: tree frog, (iii) class 64: green mamba, (iv) class 207: golden retriever, (v) class 430: basketball, (vi) class 483: castle, (vii) class 504: coffee mug, (viii) class 817: sports car, (ix) class 933: cheese burger and (x) class 947: mushroom. CFG is applied to the entire noise interval  $\sigma(t) \in [0.002, 80]$ , with guidance strength  $\gamma = 4$ . The results for CIFAR-10 and ImageNet are shown in Figure 10 and Figure 11 respectively.

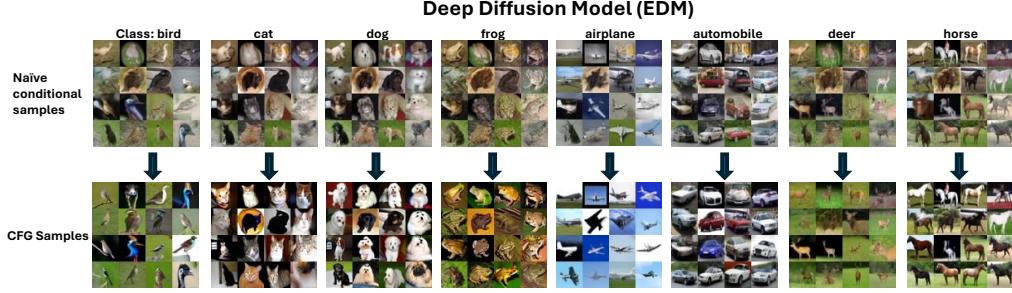
Our key observations are as follows:

- **Linear Diffusion Models.** Despite being built from class-specific means and covariances, the conditional *linear* diffusion models produce visually similar samples that lack distinguishable class features. From (8), we see that the generated samples are largely shaped by each class’s covariance structure; hence, their indistinct and low-quality generations suggest that these covariance matrices are insufficiently distinctive.
- **Deep diffusion models (EDM)** These models inherit similar limitations. The generated samples often exhibit poor image quality, with incoherent features that blend into the background and the class-specific image structures can be hard to discern. Furthermore, images generated from the same initial noise can appear structurally similar even under different class labels, indicating that naive conditional sampling fails to capture distinct, class-specific patterns. Lastly, comparing the generations from linear model and EDM reveals they match in terms of the overall structures, underscoring the key role of covariance in shaping higher-level features. Consequently, when class-specific covariance matrices are not sufficiently distinct, sample quality remains limited—even in nonlinear models.

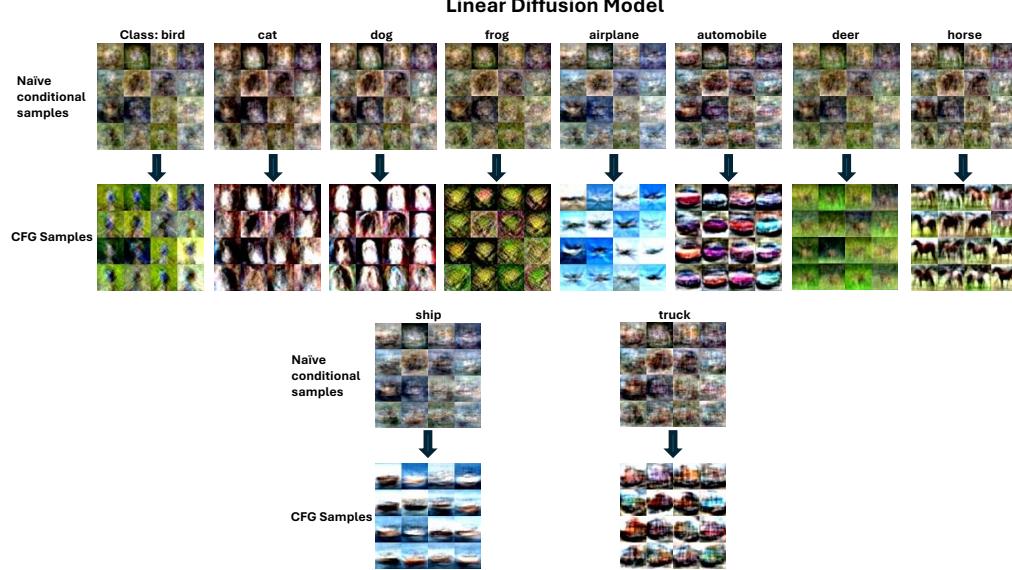
### D.2 Quantitative Results

To quantify the class-specificity gap, we compare the pairwise class similarity with FID score [27], which measures the similarity between two datasets  $X$  and  $Y$  in the Inception embedding space. For every ordered pair of different classes  $(c_i, c_j)$  we build two datasets  $(X, Y)$  and compute  $\text{FID}(X, Y)$  under three settings:

- Real data.  $X$  and  $Y$  contain all training images from classes  $c_i$  and  $c_j$ , respectively.



(a) Naïve conditional and CFG generations for EDM



(b) Naïve conditional and CFG generations for linear diffusion model

Figure 10: **Effects of CFG on CIFAR-10.** (a) and (b) demonstrate the naive conditional samples and the CFG-guided samples of deep diffusion model and linear diffusion model respectively. Each grid corresponds to the same initial noise.

- Naïve conditional sampling.  $X$  and  $Y$  contain images generated by vanilla conditional sampling (4) with the EDM model. We generate approximately the same number of images as the corresponding training images.
- Classifier-free guidance (CFG).  $X$  and  $Y$  contain images generated from the same EDM model using CFG sampling (6). We generate approximately the same number of images as the corresponding training images.

The results are presented in Figure 12, which shows that for most pairs of classes, when  $X$  and  $Y$  are built with images generated with naive conditional sampling, the FID (colored in orange) is consistently lower than when they are built with training data (colored in blue). Because lower FID indicates higher similarity, this results confirms that images produced by naive conditional sampling

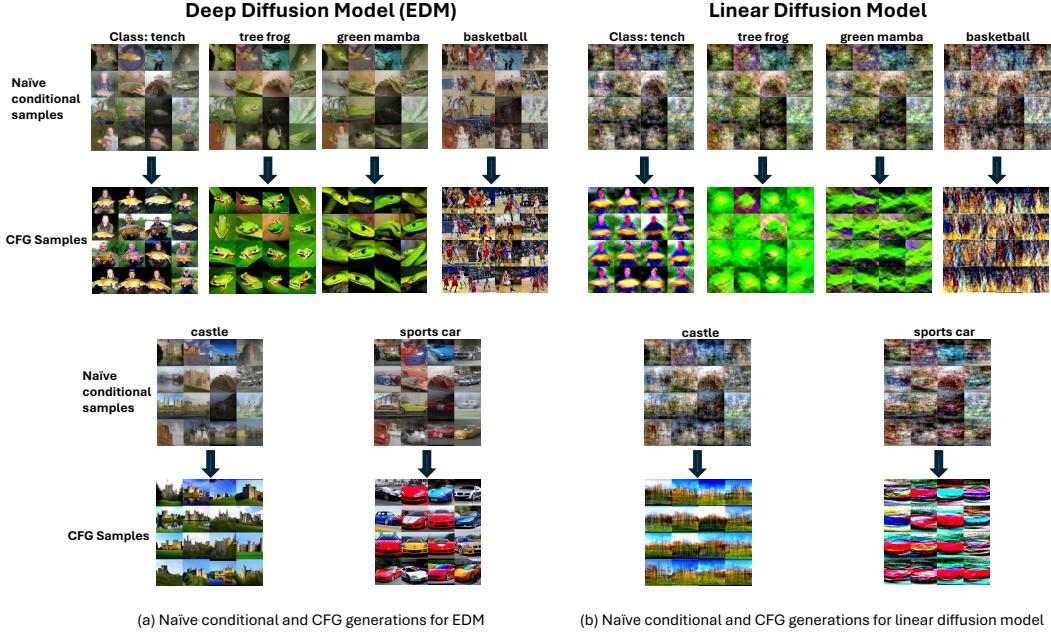


Figure 11: **Effects of CFG on ImageNet.** (a) and (b) demonstrate the naive conditional samples and the CFG-guided samples of deep diffusion model and linear diffusion model respectively. Each grid corresponds to the same initial noise. Here we only display 6 classes since the other 4 classes are presented in fig. 2.

FID between classes: training data / naïve conditional samples / CFG-guided samples											
	tench	tree frog	green mamba	golden retriever	basketball	castle	coffee mug	sports car	cheeseburger	mushroom	
tench	0	223.8 / 215.0 / 279.1	238.8 / 227.4 / 247.0	240.4 / 223.6 / 271.6	271.6 / 281.5 / 308.7	220.3 / 216.7 / 247.9	223.3 / 217.1 / 259.1	221.3 / 223.8 / 260.6	213.1 / 204.9 / 231.8	210.5 / 201.9 / 267.1	
tree frog	223.8 / 216.0 / 279.1	0	167.2 / 113.3 / 254.6	237.6 / 212.9 / 278.5	260.6 / 265.0 / 283.0	229.5 / 228.4 / 256.8	200.8 / 190.5 / 247.0	225.1 / 225.5 / 281.5	187.4 / 172.7 / 228.9	187.5 / 165.8 / 252.6	
green mamba	238.4 / 227.4 / 247.0	167.2 / 113.3 / 254.6	0	281.8 / 249.0 / 300.6	304.1 / 291.6 / 325.4	263.9 / 252.4 / 269.9	248.9 / 222.0 / 267.3	253.1 / 245.2 / 277.6	239.2 / 209.3 / 244.3	237.3 / 195.5 / 282.6	
golden retriever	240.4 / 223.6 / 271.6	237.0 / 212.9 / 278.5	281.8 / 249.0 / 300.6	0	255.9 / 242.9 / 290.0	232.9 / 215.0 / 254.3	209.5 / 185.2 / 211.9	211.3 / 216.2 / 260.0	225.5 / 203.4 / 241.1	220.6 / 180.6 / 247.1	
basketball	271.6 / 261.5 / 308.7	260.6 / 265.0 / 283.0	304.1 / 291.6 / 325.4	255.9 / 242.9 / 290.0	0	250.1 / 244.4 / 279.4	218.2 / 212.3 / 239.5	237.9 / 218.0 / 275.3	255.0 / 267.0 / 271.7	252.4 / 254.2 / 378.3	
castle	220.3 / 216.7 / 247.9	229.5 / 228.4 / 256.8	263.9 / 252.4 / 269.9	232.8 / 215.9 / 254.3	250.1 / 244.4 / 279.4	0	208.2 / 196.2 / 228.0	210.3 / 200.5 / 259.2	225.6 / 219.9 / 230.9	212.0 / 210.7 / 238.7	
coffee mug	223.3 / 217.2 / 259.1	200.8 / 190.5 / 247.0	248.9 / 222.9 / 267.2	209.5 / 185.2 / 241.9	218.2 / 212.3 / 239.5	208.2 / 196.2 / 228.0	0	188.2 / 173.8 / 225.7	178.3 / 163.5 / 219.5	197.2 / 184.9 / 240.2	
sports car	221.3 / 223.8 / 260.6	225.1 / 225.5 / 281.5	253.1 / 245.2 / 277.6	211.3 / 216.2 / 260.0	217.8 / 208.0 / 275.3	210.3 / 200.5 / 259.2	188.2 / 175.8 / 225.7	0	211.5 / 205.5 / 238.9	224.1 / 222.3 / 279.3	
cheeseburger	213.1 / 204.9 / 231.8	187.4 / 172.7 / 228.9	239.2 / 209.3 / 244.3	225.5 / 202.4 / 241.1	255.0 / 267.0 / 271.7	225.6 / 219.9 / 230.9	178.3 / 163.5 / 219.5	211.5 / 205.5 / 238.9	0	168.0 / 157.5 / 211.1	
mushroom	210.5 / 201.9 / 267.1	187.5 / 165.8 / 252.6	237.3 / 195.4 / 282.6	220.6 / 190.0 / 287.1	252.4 / 254.2 / 278.3	212.0 / 210.7 / 238.7	187.2 / 184.9 / 240.2	224.1 / 223.3 / 279.3	168.0 / 157.5 / 211.1	0	

Figure 12: **Class-to-Class Similarity (Measured with FID).** Each cell reports the FID between datasets of two classes, built with (i) training data (ii) data generated by naive conditional sampling and (iii) data generated by CFG sampling.

are less distinguishable across classes than the real data. In contrast, CFG greatly improves the FID score, implying an increased inter-class separation.

The samples used for calculating FID in Figure 12 are generated using 20 steps of Euler method (first-order sampler). Increasing the number of steps or switching to higher-order sampler only marginally narrows the gap. Table 1 shows the inter-class FID averaged over 10 selected classes as described in appendix D.1 with different sampling steps and sampler. Note that even when using 100 steps and second-order Heun samplers, the average inter-class FID is still considerably smaller compared to the training data (ground truth). Figure 13 qualitatively visualizes the samples generated from the same initial noise but different class labels. Despite conditioned on different labels, the generated images share high structural similarity. For certain classes, such as tree frog, green mamba and golden retriever, the class features are even hard to discern. In contrast, CFG greatly reduces the structural similarity, yielding images with clear, class-specific features.



Figure 13: **Naïve conditional sampling lacks class-specific features.** Figure (a) shows the samples generated with naïve conditional sampling using linear diffusion models. Figures (b)-(e) show the samples generated with naïve conditional sampling using the actual diffusion models with different steps and samplers. Figure (f) shows the generated samples with CFG guided sampling. Note that the generated images from linear models of different classes share high visually similarity, implying the covariance structures of different classes are not distinctive enough. Similar structural similarity can be observed in the samples of nonlinear diffusion models. CFG greatly alleviates this issue of lack of class-specificity, leading to images with clear class-specific features.

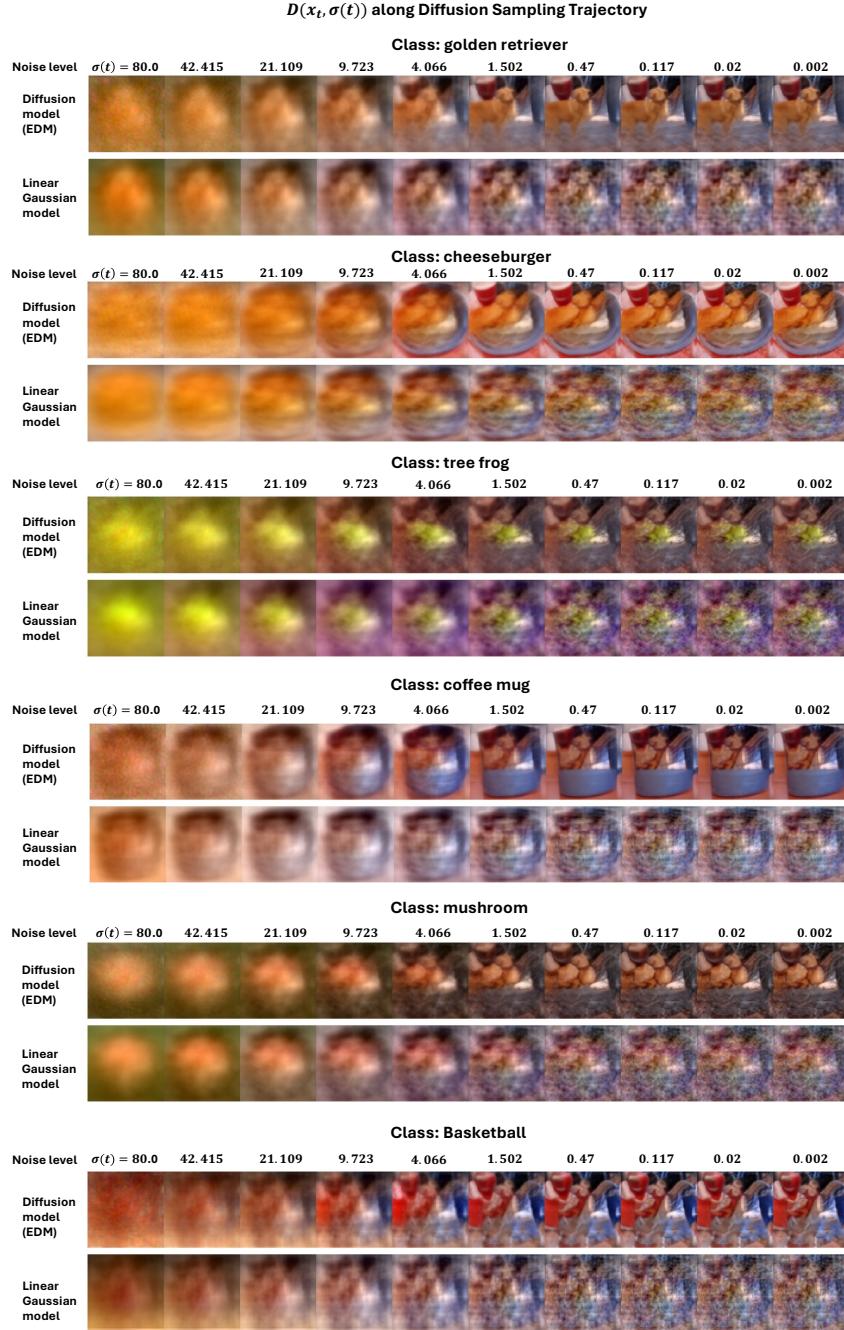


Figure 14: **Similarity between Linear and Nonlinear Models.** For high to moderate noise levels ( $\sigma(t) \in (4, 80]$ ), the linear denoisers well approximate the learned deep denoisers. Though the two models diverge in lower noise regimes, their final samples still match in overall structure. Although the linear models are built separately for each class according to (7), they generate highly similar samples when starting from the same initial noise. The same similarity also exists in the samples of real-world diffusion models.

Table 1: Average inter-class FID for training data and various sampling settings (10-class average).

Method	Steps	Sampler	Avg. FID
Training (ground truth)	—	—	226.6
Naive conditional	10	Euler	210.7
	20	Euler	214.6
	30	Euler	215.8
	50	Euler	215.9
	100	Euler	216.2
	100	Heun	216.3
	20	Euler	258.9
CFG guided ( $\gamma = 4$ )			

### D.3 Covariance Matrices of Different Classes Lack Class-Specificity

The lack of class-specificity is especially pronounced in linear diffusion models. As shown in Figure 13(a) and Figure 14, although the linear diffusion models are separately parameterized with the class-specific means and covariances for each class, the resulting samples share high similarity. Since the generated samples of the linear models are governed by the data covariances, the observed inter-class similarity implies that the covariance structures of different classes are not distinct enough.

Next, we quantitatively demonstrate that the class-specific covariance matrices are insufficiently distinct. To do this, we take  $\mathbf{U}_{uc}$ , the principal components (PCs) of the *unconditional* dataset (i.e., the singular vectors of the unconditional covariance), as a baseline. We then compare  $\mathbf{U}_{uc}$  to  $\mathbf{U}_c$ , the PCs of each *conditional* dataset. As shown in Figure 15 and Figure 16, the correlation matrices  $\mathbf{U}_c^T \mathbf{U}_{uc}$  for 10 classes (5 from CIFAR-10 and 5 from ImageNet) reveal that the leading PCs of each class share high similarity with those of the unconditional data. Thus, the PCs do not necessarily capture the distinctive features of individual classes though they represent the dominant variations of the dataset. Instead, these PCs often reflect global intensity or foreground-background variations.

**Why Covariance Structure Matters?** Covariance structures are fundamental statistics of a target distribution, and we would expect a robust diffusion model to learn them accurately. However, because these covariance structures are not sufficiently distinct, linear diffusion models—relying heavily on covariance for generation—struggle to produce high-quality images. To achieve better fidelity, models must leverage higher-order information beyond covariance. Recent works [21, 22] observe that deep diffusion models can be approximated *unreasonably* well by linear diffusion models, especially when the model capacity is limited or the training is insufficient [21]. Qualitatively, we have demonstrated the similarity between linear models and the actual diffusion models by showing that linear models replicate the coarse (low-frequency) features of samples generated by deep diffusion models. These results suggest that deep diffusion models may have an implicit bias toward learning simpler structures such as covariance, and thus the suboptimal nature of data covariance for generation task can limit their generative quality.

## E Mechanism of Linear CFG

In the setting of linear diffusion model, (37) and (38) together form the CFG guidance. For the following discussion, we let  $\tilde{\Sigma}_{c,t} = \mathbf{U}_c \tilde{\Lambda}_{\sigma(t),c} \mathbf{U}_c^T$  and  $\tilde{\Sigma}_{uc,t} = \mathbf{U}_{uc} \tilde{\Lambda}_{\sigma(t),uc} \mathbf{U}_{uc}^T$ .

### E.1 Mean-Shift Guidance

Equation (38) is the mean-shift guidance term that shifts  $\mathbf{x}_t$  towards  $(\mathbf{I} - \tilde{\Sigma}_{uc,t})(\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc})$ , a direction independent of the specific sample  $\mathbf{x}_t$ . At sufficiently large  $\sigma(t)$ ,  $(\mathbf{I} - \tilde{\Sigma}_{uc,t})(\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc}) \approx \boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc}$ , indicating the mean-shift term approximately shifts  $\mathbf{x}_t$  towards the direction of the difference between class mean and unconditional mean  $\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc}$ . As  $\sigma(t)$  decreases, the components of  $\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc}$  within the subspace spanned by the unconditional PCs ( $\mathbf{U}_{uc}$ ) progressively shrink to 0. Figure 17 demonstrates  $\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc}$  and the evolution of the mean-shift guidance term  $(\mathbf{I} - \tilde{\Sigma}_{uc,t})(\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc})$  across different noise levels. Notice that for a wide range of noise levels  $\sigma(t)$ ,  $(\mathbf{I} - \tilde{\Sigma}_{uc,t})(\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc})$

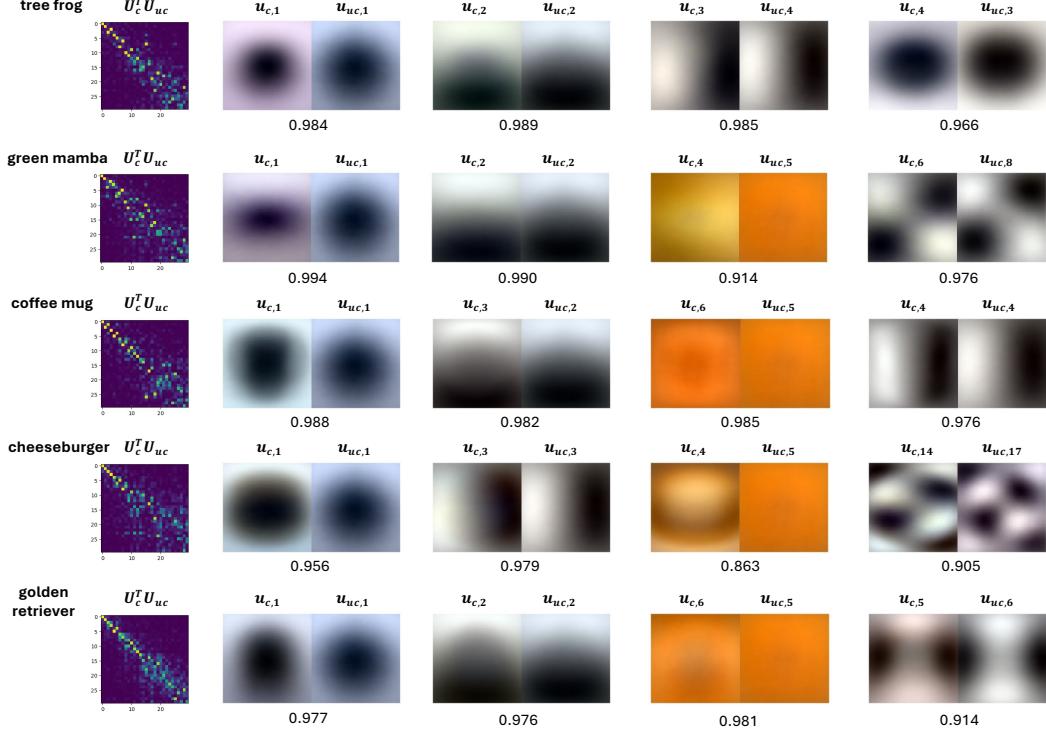


Figure 15: **Covariance Structures of CIFAR-10.** Each row corresponds to a different class. On the left, we show the correlation matrix between conditional and unconditional principal components (PCs), visualizing only the first 25. The subsequent images depict several highly correlated PCs, with correlation values displayed underneath. These results illustrate that the leading PCs do not always capture class-specific patterns.

remains close to  $\mu_c - \mu_{uc}$ , before it becomes uninformative at small  $\sigma(t)$ . Hence, as stated in the main text, the mean-shift guidance term has the effect of approximately shifting  $\mathbf{x}_t$  in the direction  $\mu_c - \mu_{uc}$ .

## E.2 CPC guidance

Equation (37) is the CPC guidance term. Let  $\mathbf{V}_{\sigma(t)} \hat{\Lambda}_{\sigma(t)} \mathbf{V}_{\sigma(t)}^T$  be the eigendecomposition of  $\tilde{\Sigma}_{c,t} - \tilde{\Sigma}_{uc,t}$ , whose eigen spectrum is demonstrated in Figure 18, the CPC guidance term can be further decomposed into the positive CPC and negative CPC guidance:

$$\frac{\gamma}{\sigma^2(t)} (\mathbf{V}_{\sigma(t),+} \hat{\Lambda}_{\sigma(t),+} \mathbf{V}_{\sigma(t),+}^T) (\mathbf{x}_t - \mu_c) dt, \quad (67)$$

$$\frac{\gamma}{\sigma^2(t)} (\mathbf{V}_{\sigma(t),-} \hat{\Lambda}_{\sigma(t),-} \mathbf{V}_{\sigma(t),-}^T) (\mathbf{x}_t - \mu_c) dt, \quad (68)$$

where  $\mathbf{V}_{\sigma(t),+}$  and  $\mathbf{V}_{\sigma(t),-}$  contain eigenvectors corresponding to positive and negative eigenvalues  $\hat{\Lambda}_{\sigma(t),+}$  and  $\hat{\Lambda}_{\sigma(t),-}$  respectively. As discussed in section 2.3,  $\tilde{\Sigma}_{c,t}$  and  $\tilde{\Sigma}_{uc,t}$  are (up to a factor  $\sigma(t)^2$ ) the conditional and unconditional posterior covariances of  $p_{\text{data}}(\mathbf{x}|\mathbf{c}) = \mathcal{N}(\mu_c, \Sigma_c)$  and  $p_{\text{data}}(\mathbf{x}) = \mathcal{N}(\mu_{uc}, \Sigma_{uc})$ . Hence,  $\mathbf{V}_{\sigma(t)}$  are the CPCs which contrast between  $X \sim p_{\text{data}}(\mathbf{x}|\mathbf{x}_t, \mathbf{c})$  and  $Y \sim p_{\text{data}}(\mathbf{x}|\mathbf{x}_t)$ . Specifically,  $\mathbf{V}_{\sigma(t),+}$  captures directions of higher conditional variance (class-specific features), while  $\mathbf{V}_{\sigma(t),-}$  captures directions of higher unconditional variance (features more relevant to the unconditional data). Figure 19 illustrates the evolution of positive CPCs ( $\mathbf{V}_{\sigma(t),+}$ ) and PCs ( $\mathbf{U}_c$ ) across different noise levels. It is evident that the positive CPCs better capture the class-specific patterns compared to PCs. Here we choose not to display negative CPCs since they correspond to generic features that explain more variances for the unconditional dataset, which

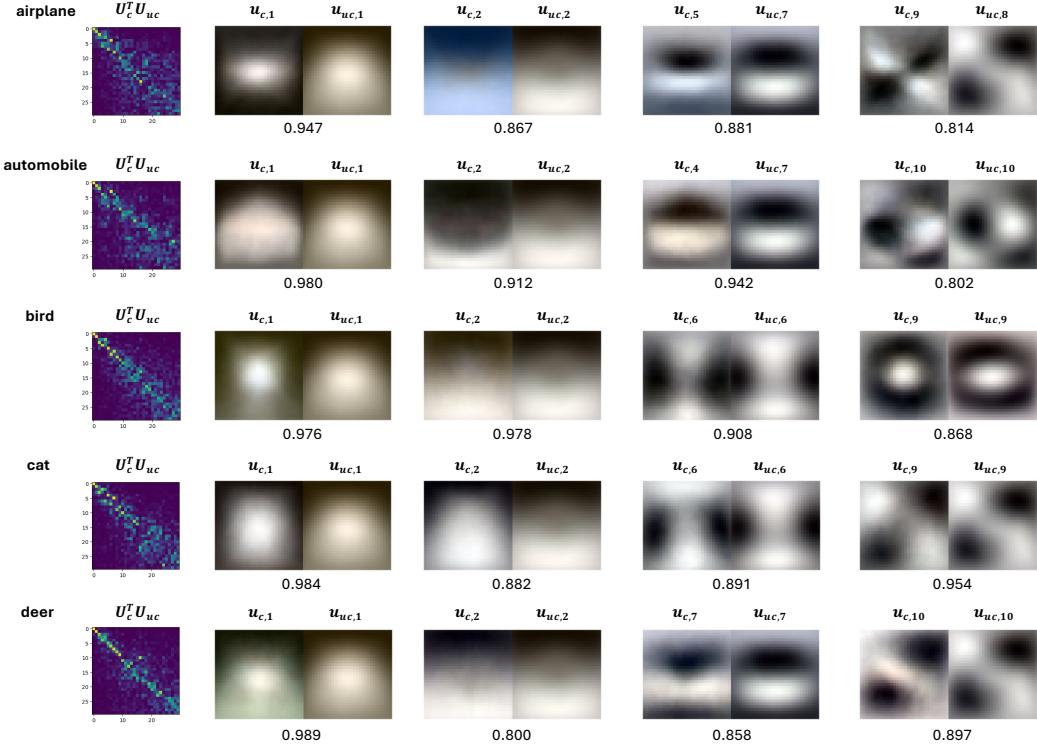


Figure 16: **Covariance Structures of ImageNet.** Each row corresponds to a different class. On the left, we show the correlation matrix between conditional and unconditional principal components (PCs), visualizing only the first 25. The subsequent images depict several highly correlated PCs, with correlation values displayed underneath. These results illustrate that the leading PCs do not always capture class-specific patterns.

are less visually interpretable. Nevertheless, as we will show next, suppressing these directions is beneficial.

### E.3 Distinct Effects of the CFG Components

As we discussed in the main text, the three CFG components have the following effects respectively:

- The positive CPC guidance term amplifies the components of  $\mathbf{x}_t$  that lie in the subspace spanned by the positive CPCs, thereby enhancing class-specific patterns.
- The negative CPC guidance term suppresses components of  $\mathbf{x}_t$  that lie in the subspace spanned by the negative CPCs, mitigating background clutter and irrelevant content. As a result, the class-relevant structures become more salient.
- The mean-shift term approximately shifts  $\mathbf{x}_t$  in the direction  $\mu_c - \mu_{uc}$ , enhancing the structure of class mean within the generated samples. However, since this perturbation is independent of the specific  $\mathbf{x}_t$ , it tends to reduce sample diversity.

**Qualitative Results.** Figures 20 and 21 qualitatively demonstrates the effects of each CFG component in linear diffusion models over 10 different ImageNet classes.

**Quantitative Results.** The distortion effects of the CFG components can be quantitatively verified through the following experiment:

- For a chosen class, generate 1,000 samples using naive conditional sampling (denote the samples as  $\mathbf{x}_c$ ) and 1,000 samples using CFG (denote the samples as  $\mathbf{x}_{cfg}$ ), and center both sets by subtracting the class mean  $\mu_c$ .

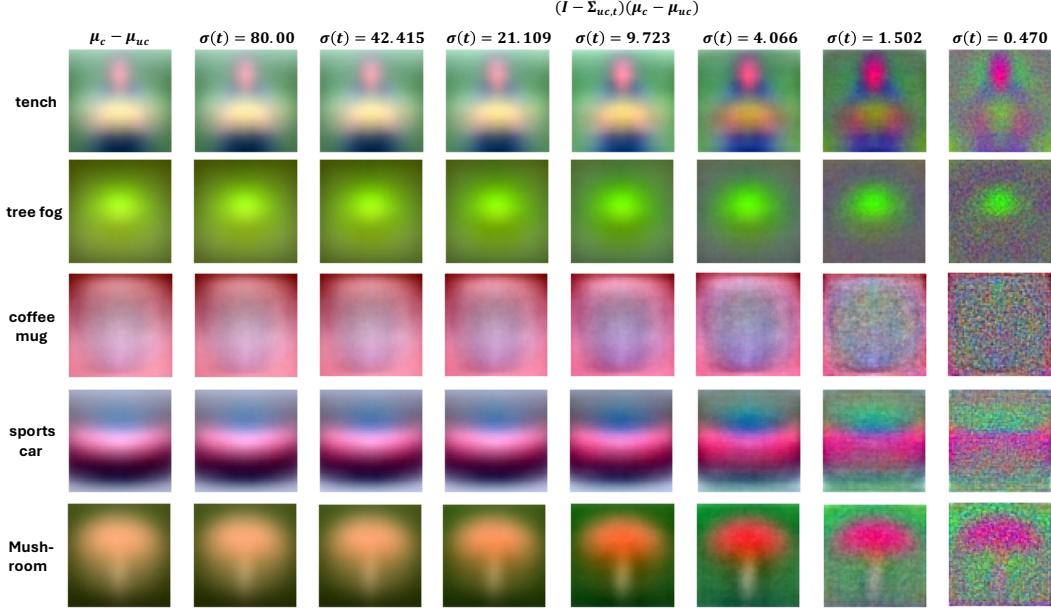


Figure 17: **Evolution of Mean-shift Guidance.** The leftmost image shows  $\mu_c - \mu_{uc}$  while the subsequent images illustrate  $(I - \tilde{\Sigma}_{uc,t})(\mu_c - \mu_{uc})$  at various noise levels  $\sigma(t)$ . Note that over a wide range of  $\sigma(t)$ ,  $(I - \tilde{\Sigma}_{uc,t})(\mu_c - \mu_{uc})$  remains close to  $\mu_c - \mu_{uc}$ .

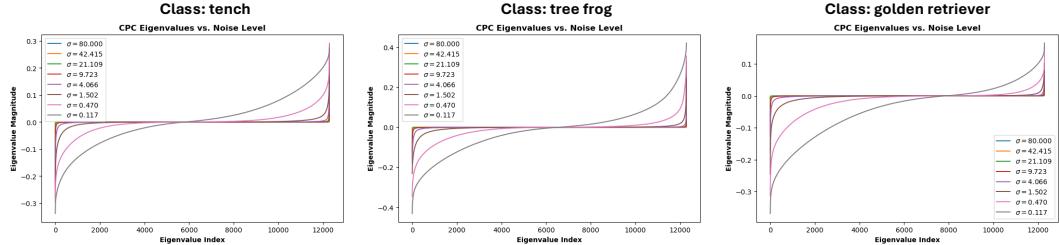


Figure 18: **Eigenvalues of  $\tilde{\Sigma}_{c,t} - \tilde{\Sigma}_{uc,t}$ .** The matrix  $\tilde{\Sigma}_{c,t} - \tilde{\Sigma}_{uc,t}$  exhibits both positive and negative eigenvalues, whose corresponding eigenvectors correspond to positive and negative CPCs respectively. Though we only show the spectrum for three classes, this behavior remains consistent across other classes.

- For a chosen positive (or negative) CPC  $v$ , compute the projection magnitudes  $|v^T(\mathbf{x}_c - \mu_c)|$  and  $|v^T(\mathbf{x}_{cfg} - \mu_c)|$  to obtain a series of univariate distributions along  $v$ .
- Project the same samples onto the mean-shift direction  $\mu_c - \mu_{uc}$  by performing  $(\mu_c - \mu_{uc})^T(\mathbf{x}_c - \mu_c)$  and  $(\mu_c - \mu_{uc})^T(\mathbf{x}_{cfg} - \mu_c)$ .

The resulting univariate distributions quantify the amount of energy the samples have along these directions. The above experiment are performed on both linear and nonlinear (EDM) diffusion models. The samples are generated using 20 steps and the guidance strength  $\gamma$  is set to 2. We focus on the first class of ImageNet (tench) and present the results on the first 5 positive CPCs and negative CPCs. As shown in Figures 22 and 23, compared to the samples with no CFG, the distributions of the CFG-guided samples have higher density on the positive CPC directions but lower density on the negative CPC directions, implying the former is enhanced while the latter is suppressed. The univariate distribution of the projection onto the mean-shift direction is presented in Figure 4(c) (bottom row), from which it is clear that the density is shifted in the direction of  $\mu_c - \mu_{uc}$ .

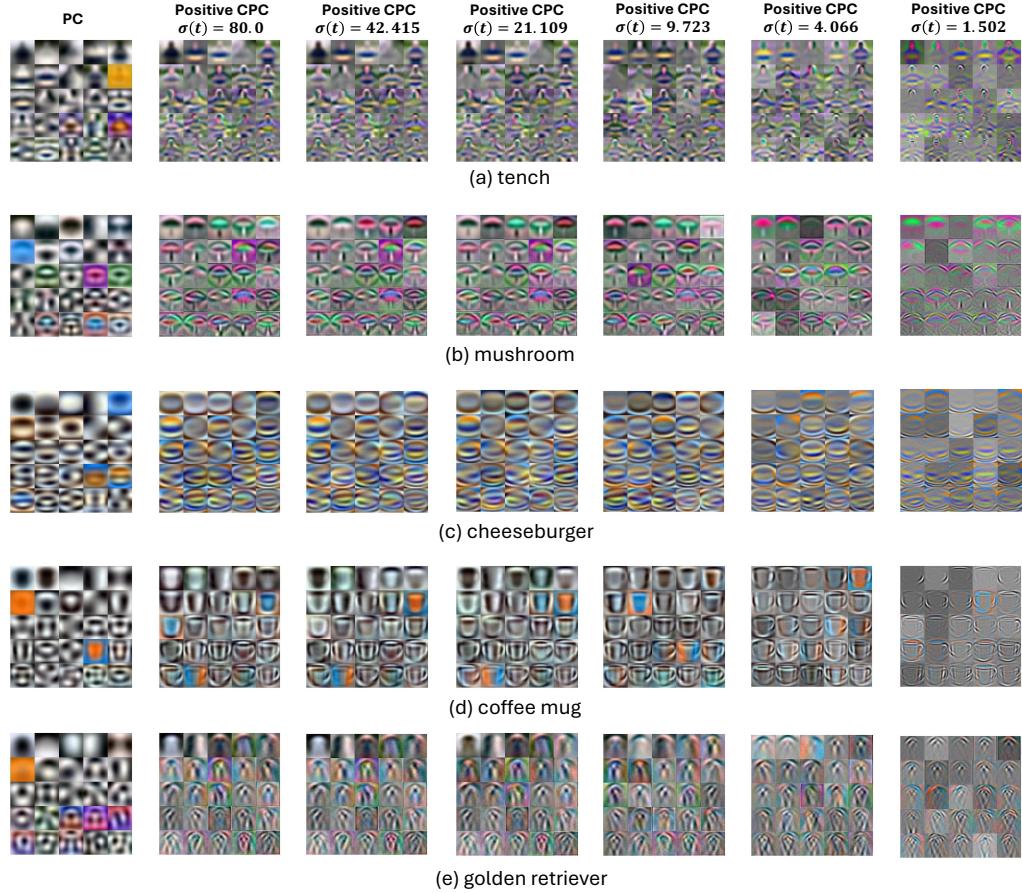


Figure 19: **Visualization of PCs and Positive CPCs.** Compared to principal components (PCs), the positive CPCs better capture class-specific patterns. Although only five classes are shown here, similar trends appear across other classes as well.

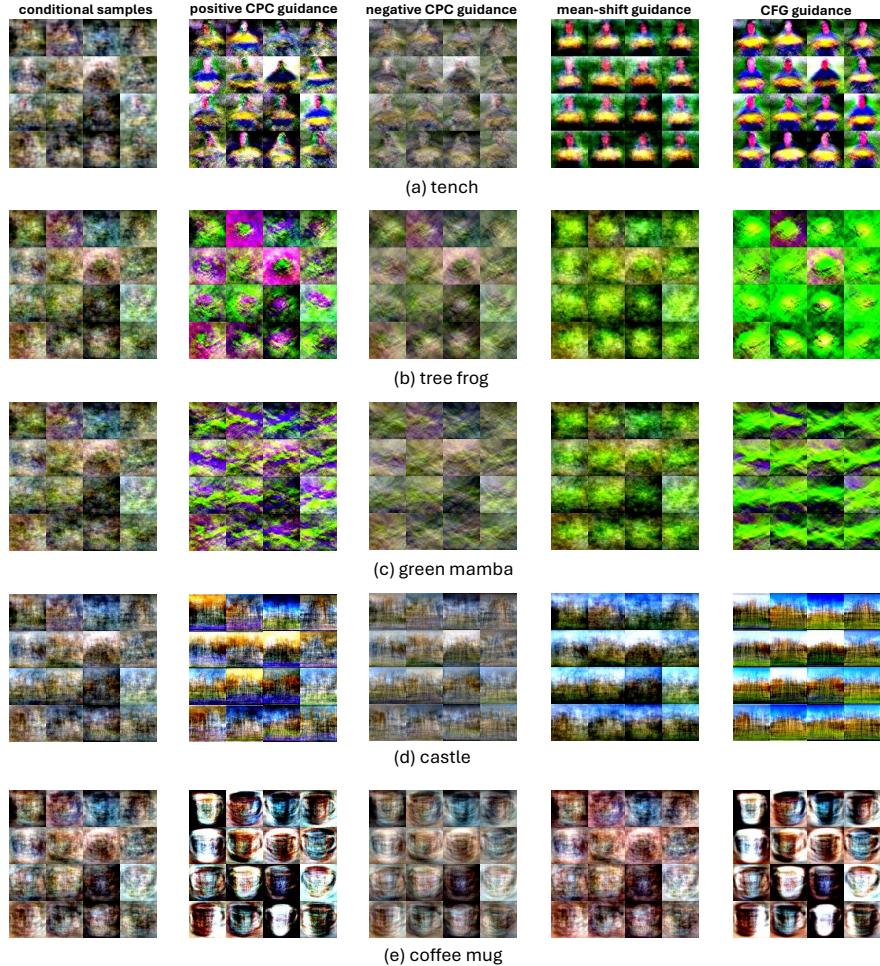


Figure 20: **Distinct Effects of Different CFG Components.** Each row shows (from left to right) the samples generated with (i) naive conditional sampling, (ii) guided with positive CPC term only (ii) guided with negative CPC term only, (iii) guided with mean-shift term only and (iv) guided with the full complete CFG. Each row corresponds to a different class.

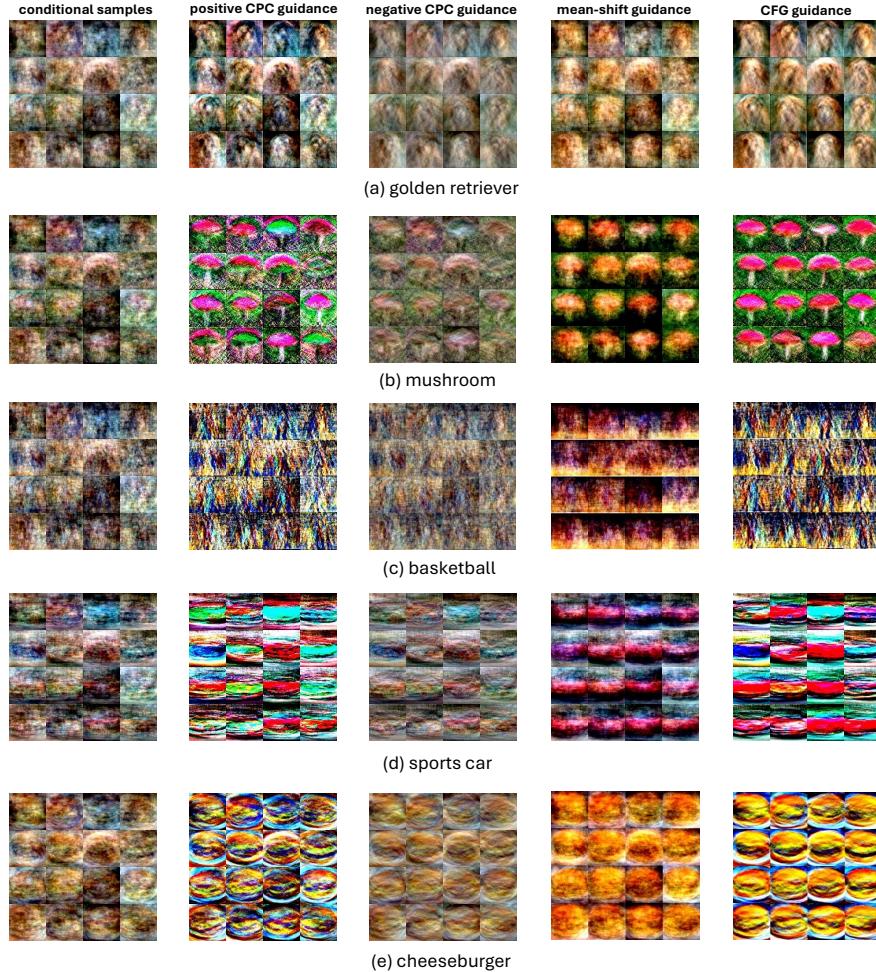


Figure 21: **Distinct Effects of Different CFG Components.** Each row shows (from left to right) the samples generated with (i) naive conditional sampling, (ii) guided with positive CPC term only (ii) guided with negative CPC term only, (iii) guided with mean-shift term only and (iv) guided with the full complete CFG. Each row corresponds to a different class.

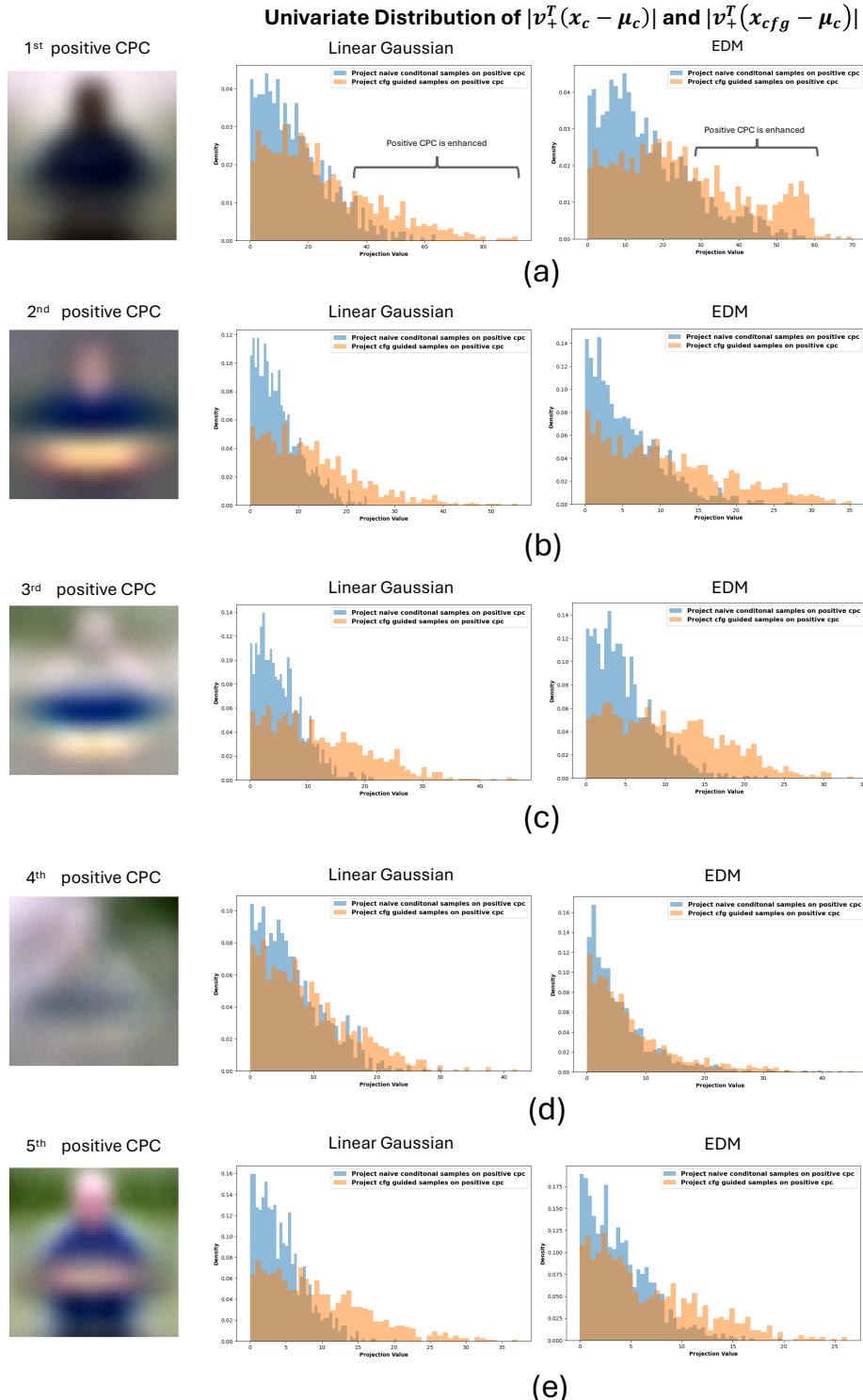


Figure 22: **CFG enhances positive CPCs.** For both linear and deep diffusion models, we randomly generate 1,000 naive conditional samples  $x_c$  and CFG-guided samples  $x_{cfg}$ , center them by subtracting the class mean  $\mu_c$ , and project them onto the top 5 positive CPCs ( $v_+$ ) to obtain a series univariate distributions. In both model types, the distributions of CFG-guided samples have greater density at higher projection values, suggesting that CFG amplifies the positive CPCs.

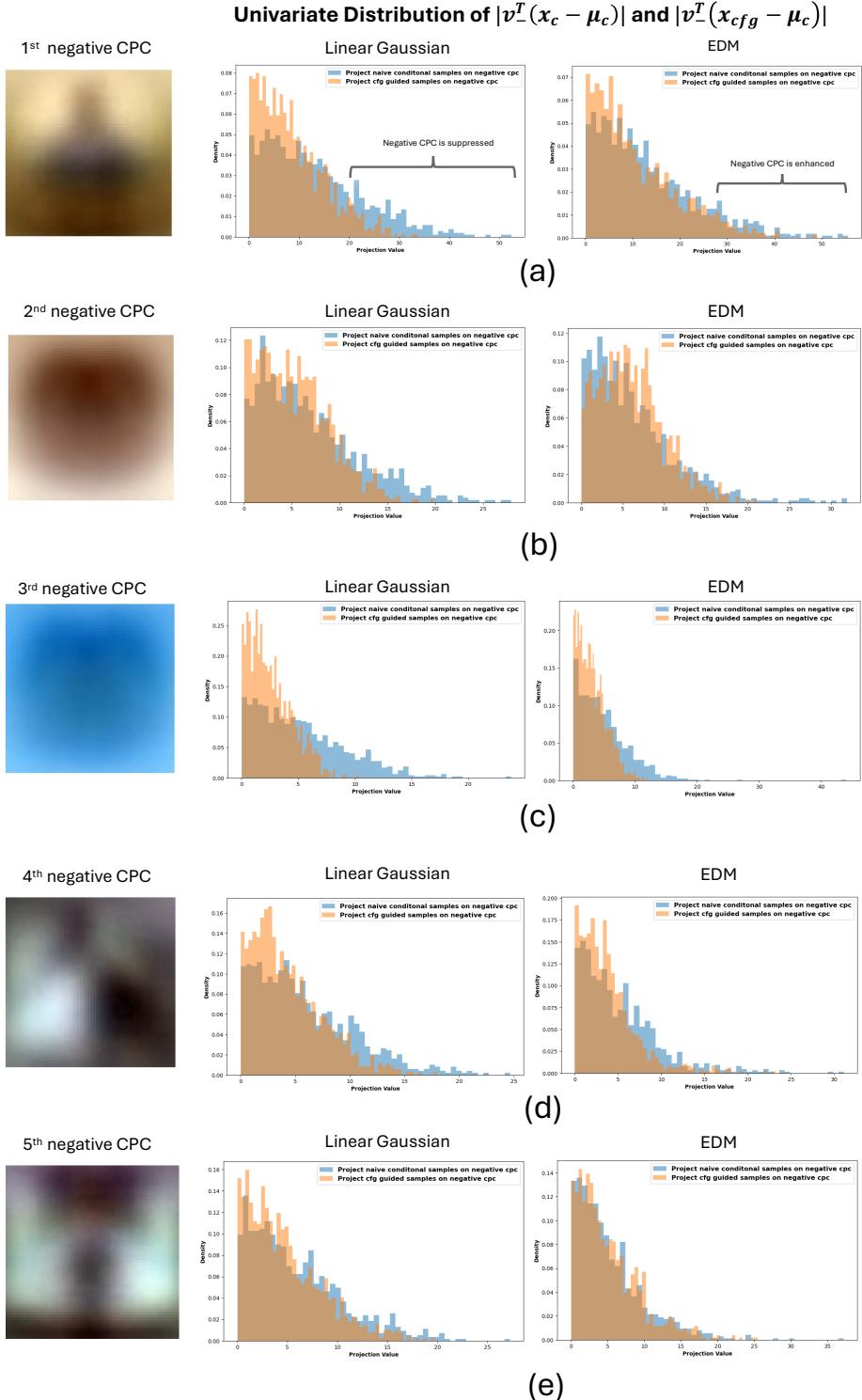


Figure 23: **CFG suppresses negative CPCs.** For both linear and deep diffusion models, we randomly generate 1,000 naive conditional samples  $x_c$  and CFG-guided samples  $x_{cfg}$ , center them by subtracting the class mean  $\mu_c$ , and project them onto the top 5 negative CPCs ( $v_-$ ) to obtain a series univariate distributions. In both model types, the distributions of CFG-guided samples have greater density at lower projection values, indicating that CFG suppresses the negative CPCs.

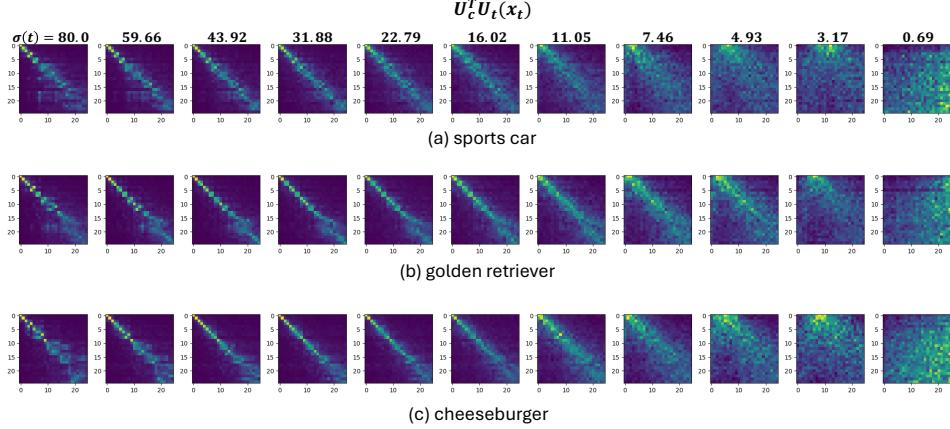


Figure 24: **Correlation between  $U_t(x_t)$  and  $U_c$ .** The leading singular vectors of  $\nabla \mathcal{D}_\theta(x_t; \sigma(t), \mathbf{c})$  well align with  $U_c$  for high to moderate  $\sigma(t)$ . Each plot shows the average correlation computed over 10 randomly initialized sampling trajectories measured for three different classes.

## F CFG in Nonlinear Deep Diffusion Models

In this section we provide additional experimental results for section 4, where we investigate how well the insights derived from linear diffusion models extend to real-world, nonlinear deep diffusion models. In this work, we study the state-of-the-art EDM models [4].

### F.1 Linear to Nonlinear Transition in Diffusion Models

Recent studies [21, 22] observe that at high to moderate noise levels, deep diffusion models  $\mathcal{D}_\theta(x_t; \sigma(t))$  can be well approximated by the corresponding linear diffusion models  $\mathcal{D}_L(x_t; \sigma(t))$  defined in (7). As the noise level decreases,  $\mathcal{D}_\theta(x_t; \sigma(t))$  becomes nonlinear. We verify this transition by the following experiment:

Let  $U_t(x_t)$  be the left singular vectors of the network Jacobians  $\nabla \mathcal{D}_\theta(x_t; \sigma(t), \mathbf{c})$  along the sampling trajectories, and let  $U_c$  be the left singular vectors of  $\nabla \mathcal{D}_L(x_t; \sigma(t), \mathbf{c})$ . Since  $\mathcal{D}_L(x; \sigma(t)) = \mu_c + U_c \tilde{\Lambda}_{c, \sigma(t)} U_c^T (x - \mu_c)$ , if  $\mathcal{D}_\theta \approx \mathcal{D}_L$ , then  $\nabla \mathcal{D}_\theta(x_t; \sigma(t), \mathbf{c}) \approx \nabla \mathcal{D}_L(x_t; \sigma(t), \mathbf{c}) \approx U_c \tilde{\Lambda}_{c, \sigma(t)} U_c^T$ , implying  $U_t(x_t) \approx U_c$ , independent of  $x_t$ . As illustrated in Figure 24, for large  $\sigma(t)$ , the leading singular vectors of  $\nabla \mathcal{D}_\theta(x_t; \sigma(t), \mathbf{c})$  indeed align with  $U_c$ . Note that since  $\tilde{\Lambda}_{c, \sigma(t)} = \text{diag}(\frac{\lambda_{c,1}}{\lambda_{c,1} + \sigma^2(t)}, \dots, \frac{\lambda_{c,d}}{\lambda_{c,d} + \sigma^2(t)})$ ,  $\nabla \mathcal{D}_L(x; \sigma(t))$  is highly low-rank at large  $\sigma(t)$ . Thus, our primary interest is in the leading singular vectors, and the non-leading singular vectors are ambiguous. In contrast, for small  $\sigma(t)$ , the alignment no longer holds and  $\nabla \mathcal{D}_\theta(x_t; \sigma(t))$  starts to adapt to individual samples, reflecting the model’s nonlinear behavior. Figures 5, 25 and 26 qualitatively demonstrates this linear to nonlinear transition.

### F.2 CFG in the Linear Regime

We provide additional experimental results for section 4.1 in Figures 27 to 30. Because the precise transition time from the linear to the nonlinear regime—as well as the influence of each CFG component—varies across classes, we empirically choose the interval for applying guidance and calculate the FD<sub>DINOv2</sub> score with 50,000 generated images for each class separately. We summarize our observations as follows (see also section 4.1):

- **Linear vs. Nonlinear CFG.** Applying linear CFG to deep diffusion models produces effects that closely resemble those of the actual (nonlinear) CFG.
- **Dominance of Mean-Shift.** In most of the 10 classes studied, the mean-shift guidance term dominates CFG behavior, as it alone can generate results visually similar to full CFG. However, for the coffee mug class, the positive CPC term takes precedence, becoming the primary driver of CFG.

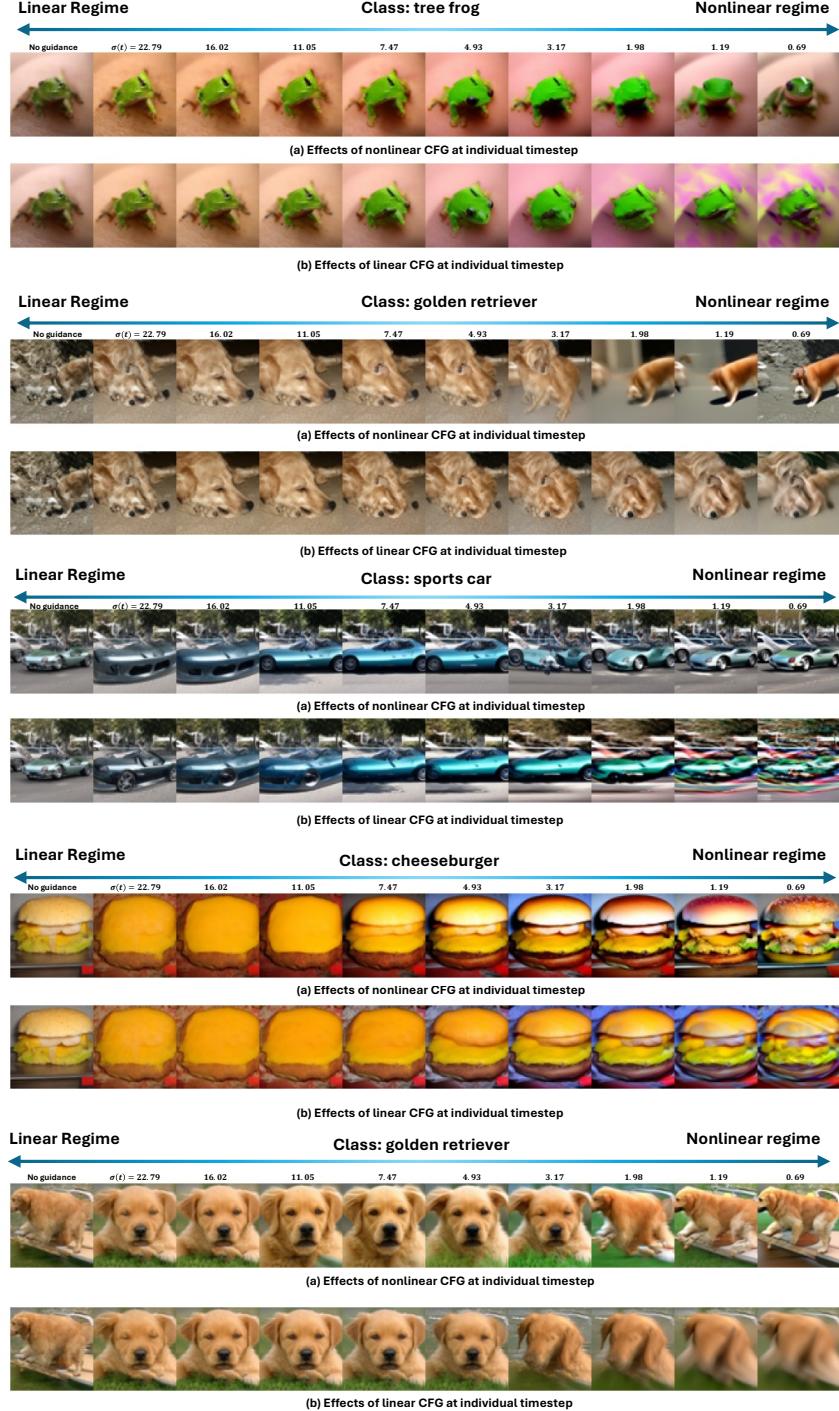


Figure 25: **Linear-to-nonlinear transition in diffusion models.** (a) and (b) compare nonlinear CFG and linear CFG applied to a deep diffusion model (EDM). The leftmost column shows unguided samples; subsequent columns show final samples when guidance is applied only at a specific noise level, with  $\gamma = 15$ .

- **Role of CPC Guidance.** CPC guidance generally improves generation quality, though its benefits can sometimes be less pronounced. For instance, in the tree frog and castle classes (Figure 29), the CPC term does not enhance  $FD_{DINO2}$  as much as the mean-shift term. Nevertheless, CPC guidance

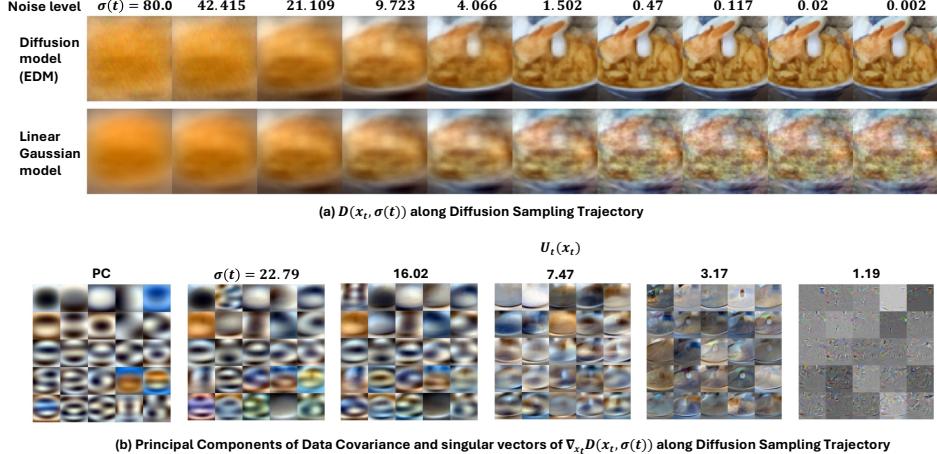


Figure 26: **Evolution of Denoiser Jacobian During Sampling.** (a) demonstrates one reverse diffusion trajectory. The left most image of (b) demonstrates the leading PCs of the data covariance. The subsequent images visualize the singular vectors,  $U_t(x_t)$ , of the denoiser Jacobian at different noise levels. Note that at early timesteps  $U_t(x_t)$  match the PCs but gradually adapt to the geometry of the sample  $x_t$ .

operates effectively over a wider range of guidance strengths  $\gamma$  and noise intervals. For the green mamba and basketball classes, we show results within the prescribed noise interval as solid curves, and extend beyond this interval as dashed curves. While mean-shift becomes highly detrimental once outside the linear regime, CPC guidance remains beneficial.

### F.3 Mean-Shifted Noise Initialization

The observation that the sample-independent mean-shift guidance alone leads to improved  $FD_{DINOv2}$  score implies that simply initializing the sampling process from a mean-shifted Gaussian,  $x_T \sim \mathcal{N}(\gamma(\mu_c - \mu_{uc}), \sigma^2(T)\mathbf{I})$ , with no additional guidance applied, can improve the generation quality, which we verify through the following experiment:

- For a chosen class and a positive scalar  $\gamma$ , generate 50,000 samples via naive conditional sampling initialized from a mean-shifted Gaussian  $\mathcal{N}(\gamma(\mu_c - \mu_{uc}), \sigma^2(T)\mathbf{I})$ . Then evaluate the sample quality with FID and  $FD_{DINOv2}$  scores.
- Repeat the above procedure across several classes and a range of  $\gamma$  values.

We perform the above experiments on 5 classes, where  $\sigma(T)$  is set to 31.9. The results are shown in Figures 31 to 35. Note that the sample quality improves with a properly chosen  $\gamma$ .

### F.4 CFG in the Nonlinear Regime

We provide additional experimental results for section 4.2 in Figure 36. We argue that effective guidance in this regime should satisfy two key properties:

- **Capture local structure of a specific sample.** As shown in Figure 25, when  $\sigma(t)$  is small, the model diverges considerably from its linear approximation, and linear CFG deviates from the actual nonlinear CFG. In this regime, CFG does not alter the overall image structure but instead refines existing features to produce crisper images. Consequently, effective guidance must adapt to each specific sample. We propose that such guidance can be derived from the network Jacobians  $\nabla \mathcal{D}_\theta(x_t; \sigma(t), c)$  evaluated at  $x_t$ . Prior work [33] shows that the singular vectors of these Jacobians, which are equivalent to the posterior covariances, adapt to the input  $x_t$ .
- **Capture class-specific patterns.** As in the linear case, the guidance must also capture class-specific patterns. This can be achieved by contrasting the conditional Jacobian  $\nabla \mathcal{D}_\theta(x_t; \sigma(t), c)$  with the

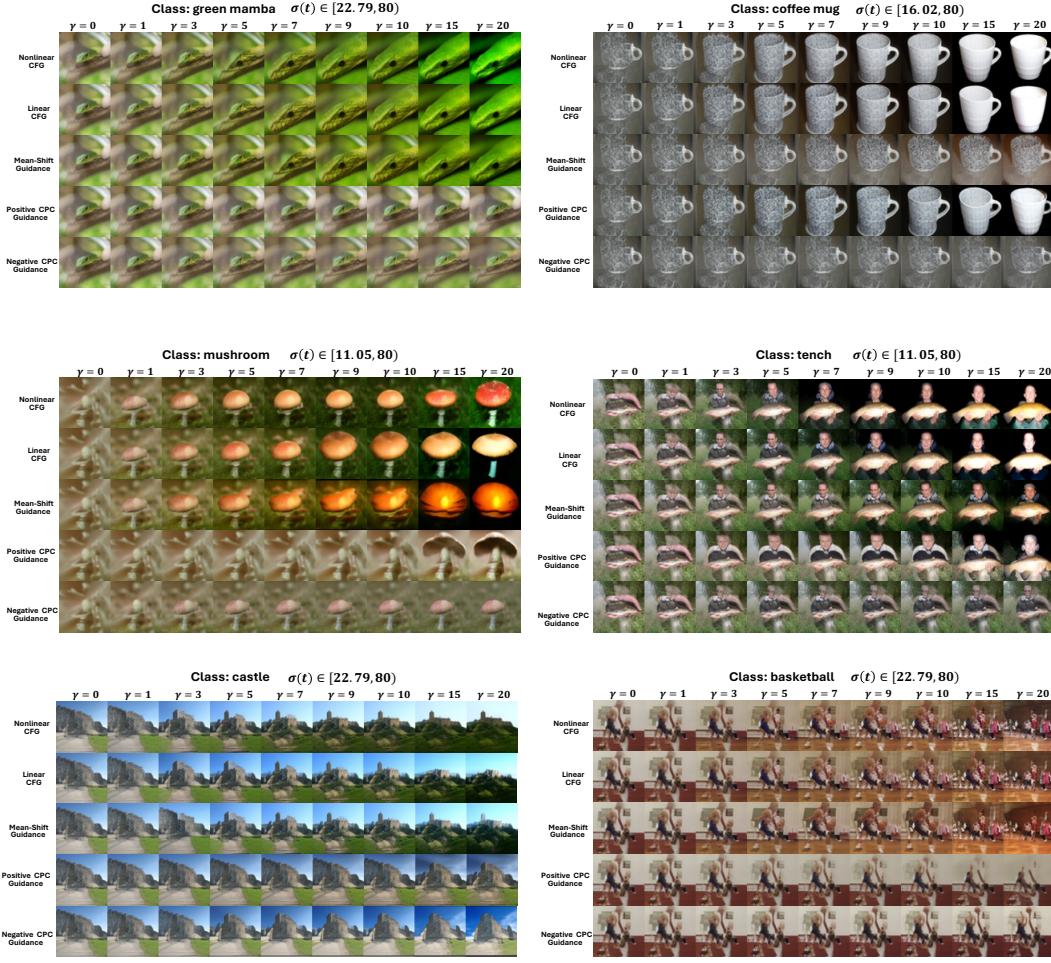


Figure 27: **Effects of CFG in Linear Regime.** Each row demonstrates the impact of different guidance types applied to EDM within the linear regime, with varying guidance strength  $\gamma$ . The guidance is applied only within intervals specified in the subtitles, where the model exhibits linear behavior.

unconditional Jacobian  $\nabla \mathcal{D}_{\theta}(\mathbf{x}_t; \sigma(t))$ . Figure 36 shows that guidance built using CPCs—i.e., the difference between these two Jacobians—yields effects similar to actual CFG. In contrast, guidance derived solely from the conditional Jacobian does not improve image quality.

Note that (17) is inspired by linear positive CPC guidance (13). We also test other guidance such as

$$\frac{\gamma}{\sigma^2(t)} \sum_i \hat{\lambda}_{+,i} \mathbf{v}_{+,i} (\mathbf{v}_{+,i}^T (\mathbf{x}_t - \boldsymbol{\mu}_c)), \quad (69)$$

but find it less effective than (17), likely due to additional noise in  $\mathbf{x}_t$ . Moreover, we observe that negative CPCs and mean-shift terms are not as effective in the nonlinear regime.

Lastly, we’d like to remark that our goal here is *not* to suggest that CFG in the nonlinear regime is exactly equivalent to (17); rather, we note that both approaches exhibit similar behaviors, implying they may share a core mechanism: identifying and amplifying *sample-specific* and *class-specific* features. The exact analytical form of CFG in the nonlinear setting remains challenging to derive due to the complexity of deep networks, leaving a promising direction for future work.

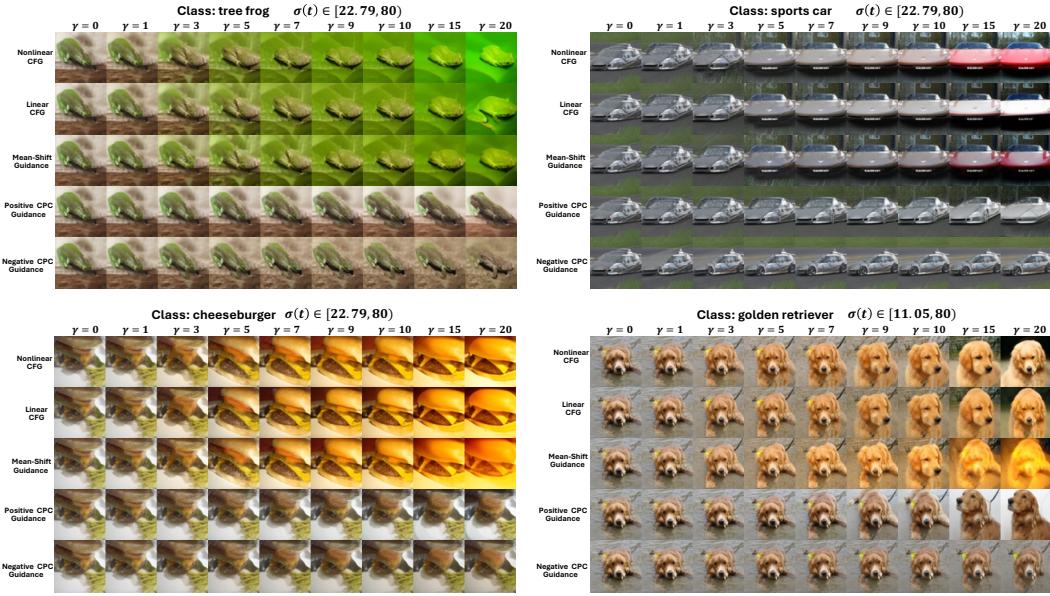


Figure 28: **Effects of CFG in Linear Regime.** Each row demonstrates the impact of different guidance types applied to EDM within the linear regime, with varying guidance strength  $\gamma$ . The guidance is applied only within intervals specified in the subtitles, where the model exhibits linear behavior.

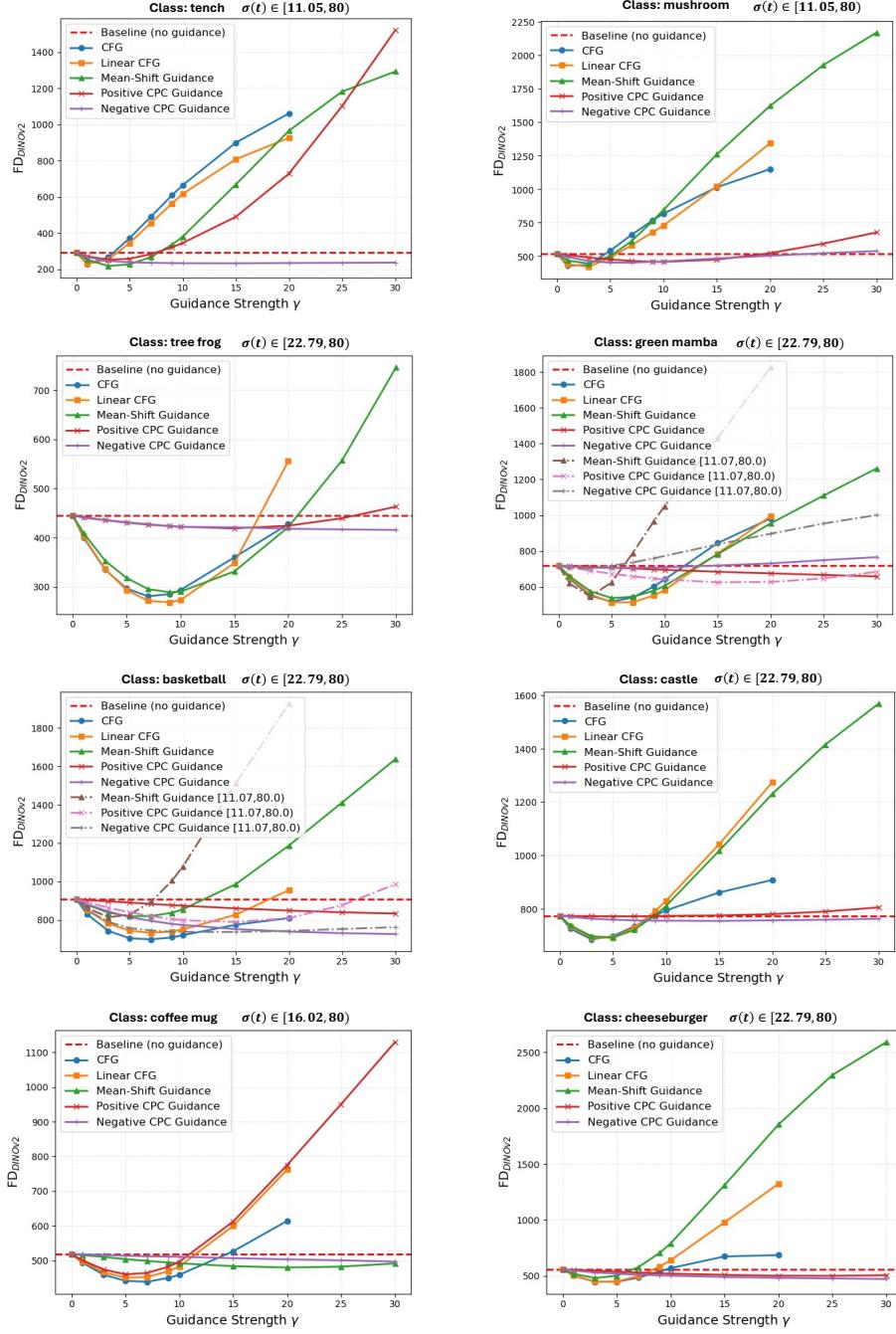


Figure 29: **FD<sub>DINOv2</sub> Scores.** The guidance is applied to the interval specified in the subtitles. For green mamba and basketball, we find it beneficial to apply CPC guidance beyond the linear regime, with results demonstrated by the dashed curves.

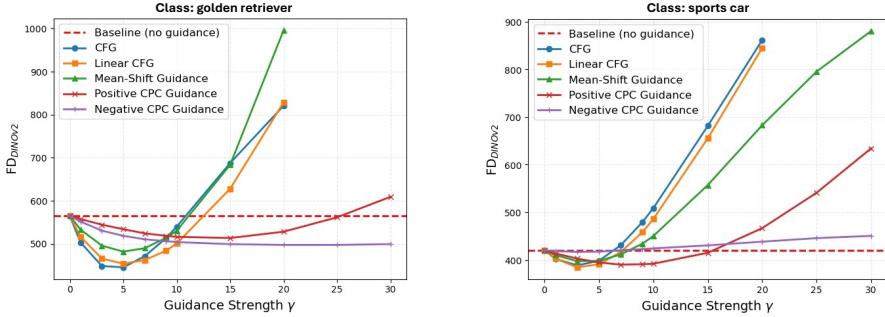


Figure 30: **FD<sub>DINOv2</sub> Scores.** The reported values are relatively high because the scores are computed separately per class, which often has a limited number of training images. It is well known that FD<sub>DINOv2</sub> scores can appear inflated when the reference dataset size is small.

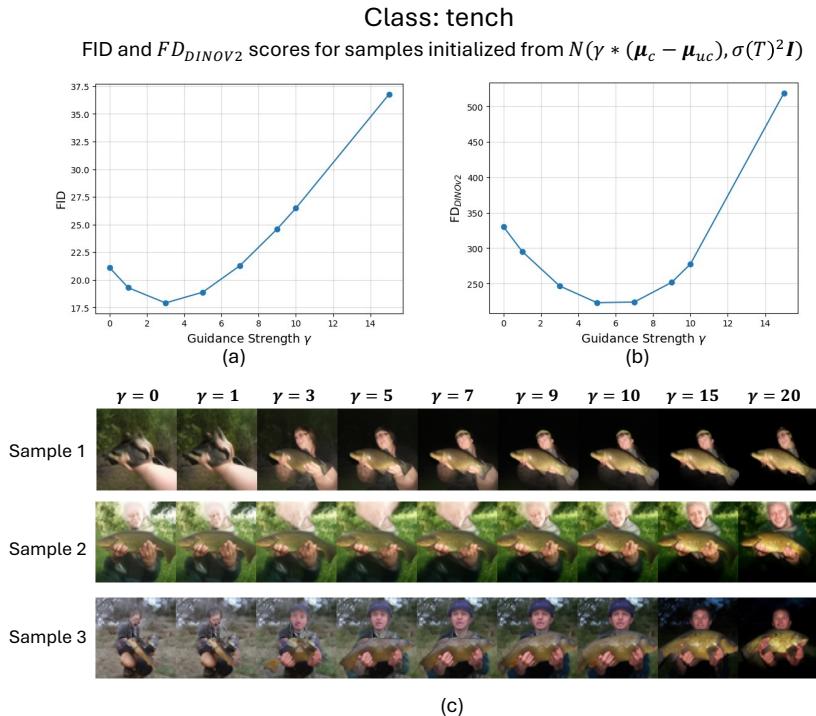


Figure 31: **Effects of initializing with mean-shift.** For every  $\gamma \in [0, 1, 3, 5, 7, 9, 10, 15, 20]$ , we generate 50,000 images from initial nosies sampled from mean-shifted Gaussian distribution  $N(\gamma(\mu_c - \mu_{uc}), \sigma(T)^2 I)$  and compute the FID scores (a) and FD<sub>DINOv2</sub> scores (b). The samples are visualized in (c). Note that adding mean-shift to the initial distribution leads to improvement of standard metrics.

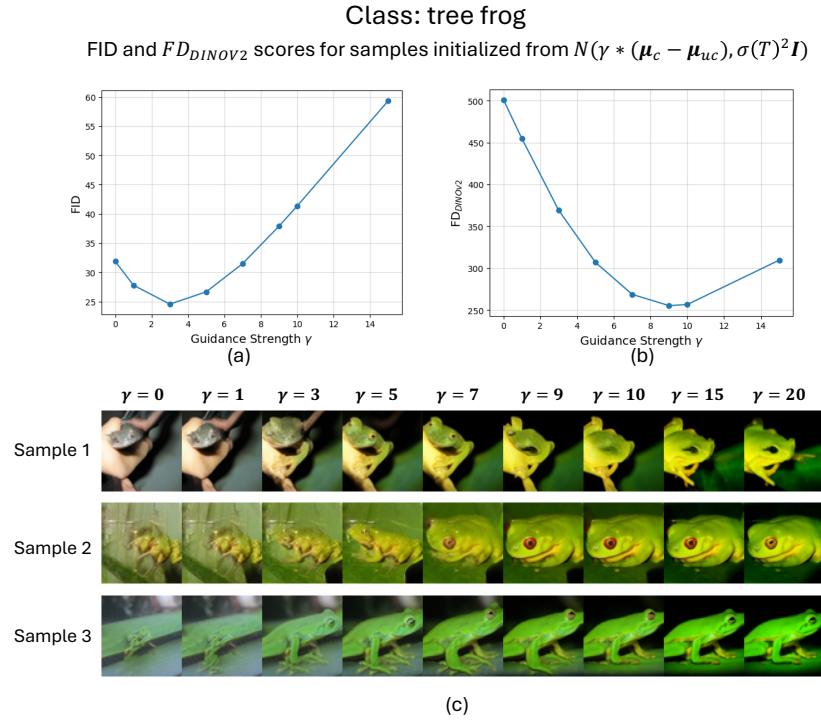


Figure 32: **Effects of initializing with mean-shift.** For every  $\gamma \in [0, 1, 3, 5, 7, 9, 10, 15, 20]$ , we generate 50,000 images from initial nosies sampled from mean-shifted Gaussian distribution  $N(\gamma(\mu_c - \mu_{uc}), \sigma(T)^2 I)$  and compute the FID scores (a) and  $FD_{DINOV2}$  scores (b). The samples are visualized in (c). Note that adding mean-shift to the initial distribution leads to improvement of standard metrics.

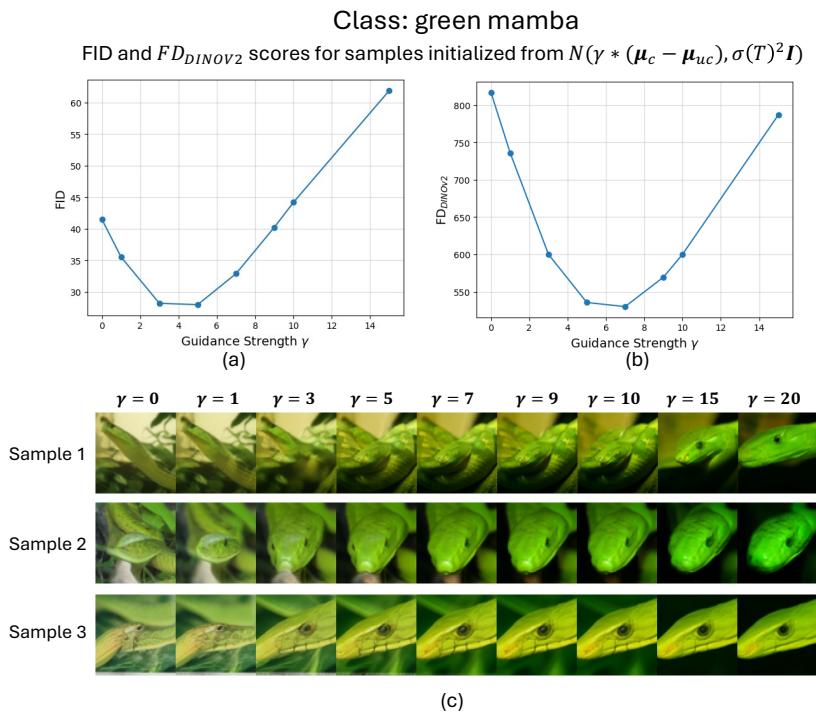


Figure 33: **Effects of initializing with mean-shift.** For every  $\gamma \in [0, 1, 3, 5, 7, 9, 10, 15, 20]$ , we generate 50,000 images from initial nosies sampled from mean-shifted Gaussian distribution  $N(\gamma(\mu_c - \mu_{uc}), \sigma(T)^2 \mathbf{I})$  and compute the FID scores (a) and  $FD_{DINOV2}$  scores (b). The samples are visualized in (c). Note that adding mean-shift to the initial distribution leads to improvement of standard metrics.

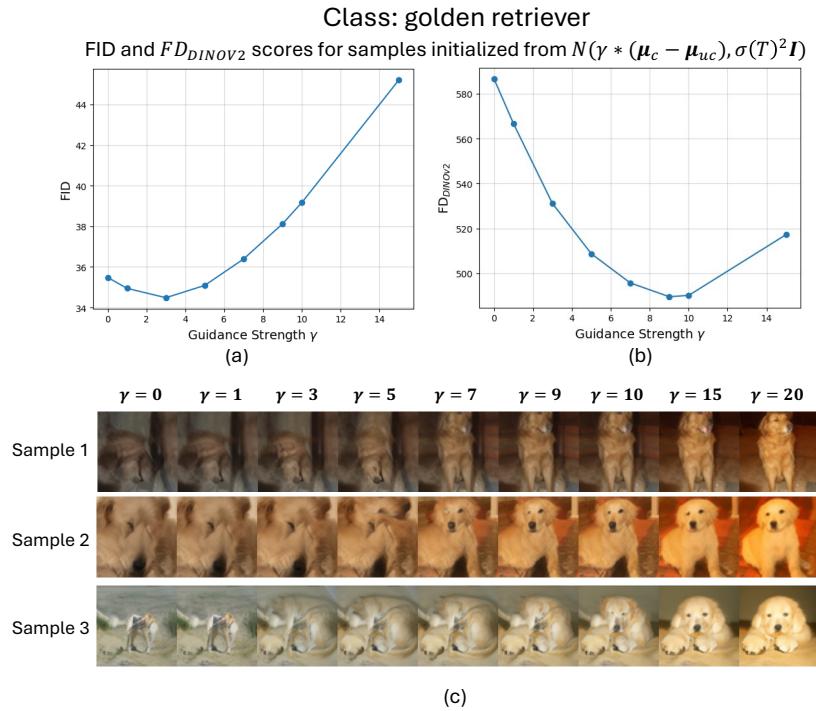
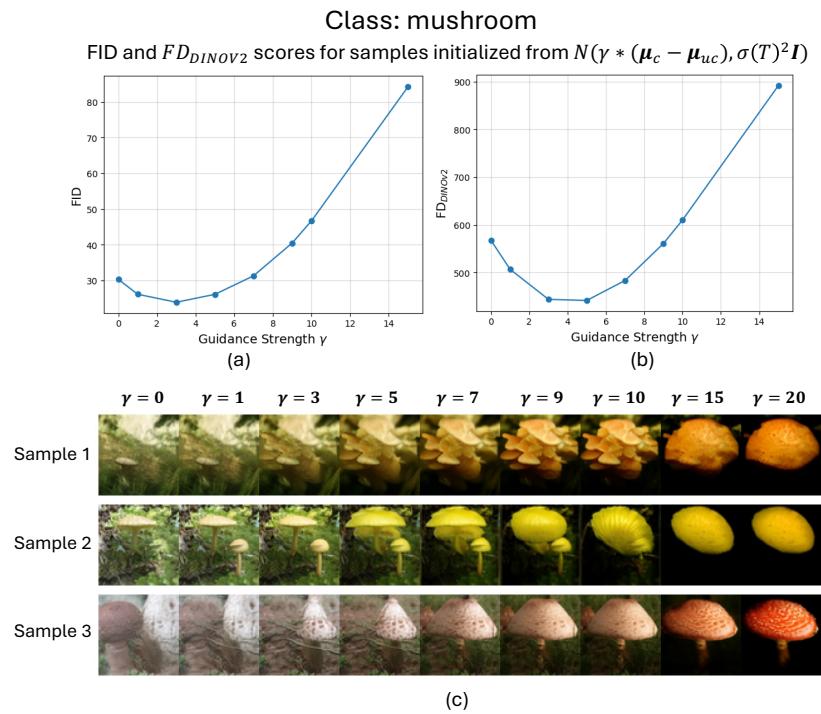


Figure 34: **Effects of initializing with mean-shift.** For every  $\gamma \in [0, 1, 3, 5, 7, 9, 10, 15, 20]$ , we generate 50,000 images from initial nosies sampled from mean-shifted Gaussian distribution  $N(\gamma(\mu_c - \mu_{uc}), \sigma(T)^2 I)$  and compute the FID scores (a) and  $FD_{DINO2}$  scores (b). The samples are visualized in (c). Note that adding mean-shift to the initial distribution leads to improvement of standard metrics.



**Figure 35: Effects of initializing with mean-shift.** For every  $\gamma \in [0, 1, 3, 5, 7, 9, 10, 15, 20]$ , we generate 50,000 images from initial nosies sampled from mean-shifted Gaussian distribution  $N(\gamma(\mu_c - \mu_{uc}), \sigma(T)^2 I)$  and compute the FID scores (a) and  $FD_{DINOV2}$  scores (b). The samples are visualized in (c). Note that adding mean-shift to the initial distribution leads to improvement of standard metrics.

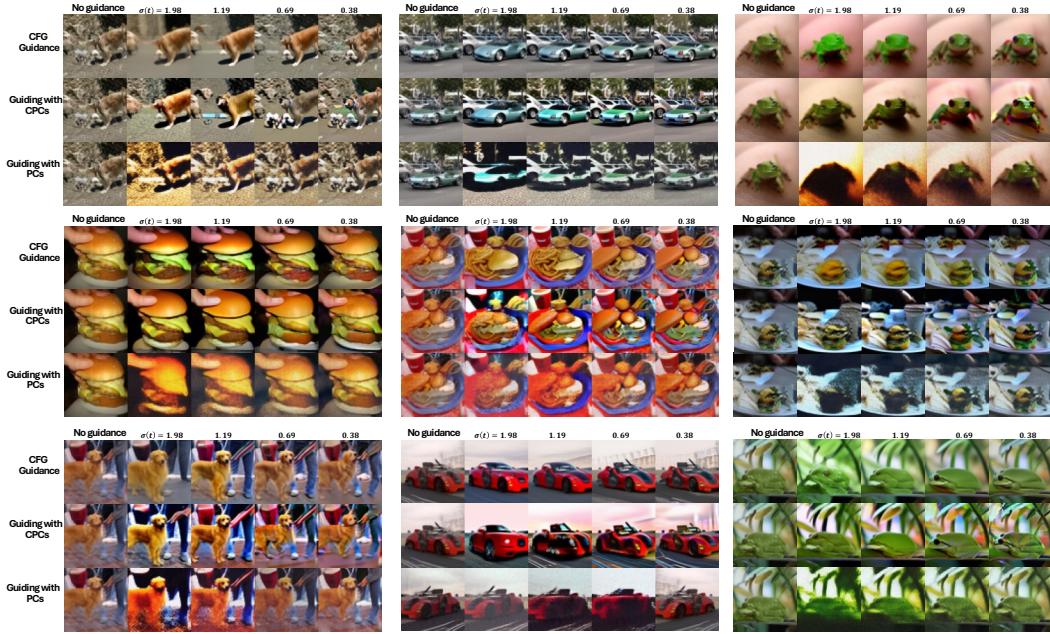


Figure 36: **Effects of CFG in the Nonlinear Regime.** Different guidance methods, each with a fixed strength of  $\gamma = 15$ , are applied at individual timesteps in the nonlinear regime. Each image shows the final output when guidance is applied solely at the timestep indicated at the top. Note that (17) closely matches the effects of CFG by enhancing finer image details, whereas (18) does not improve generation quality.

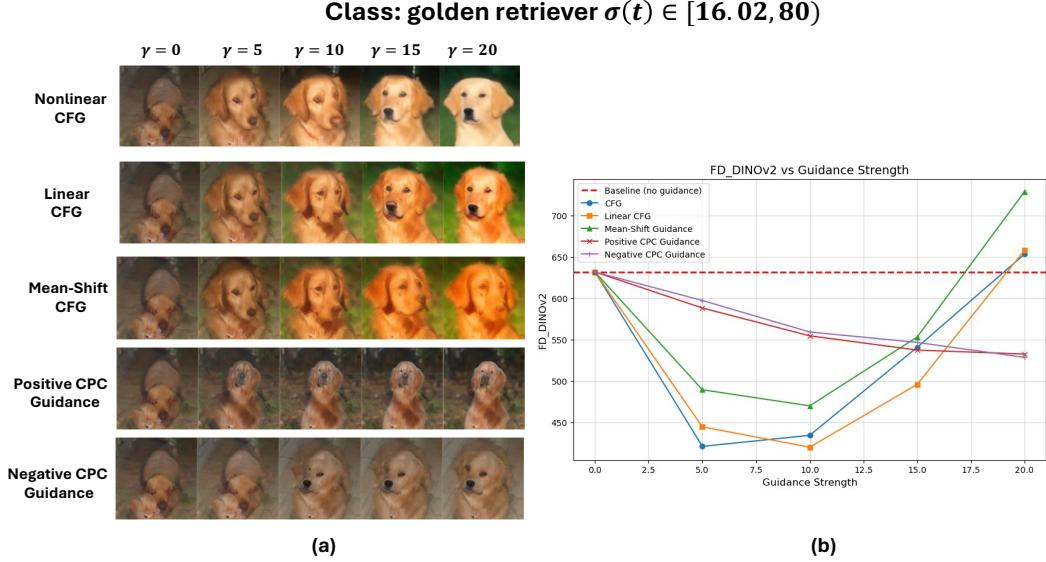


Figure 37: **Effects of CFG in the Linear Regime (EDM-2).** Each row in (a) demonstrates the impact of different guidance types applied to EDM-2 within the linear regime (specified in the subtitles), with varying guidance strength  $\gamma$ . (b) shows the FD<sub>DINOv2</sub> scores computed over 50,000 samples.

## G Experimental Results on Latent Diffusion Models

In the main text, we conducted experiments using the EDM-1 model [4], which operates directly in pixel space with  $64 \times 64$  resolution. Here, we present complementary results on the EDM-2 [5] latent diffusion model, which generates images at  $512 \times 512$  resolution.

**Linear Regime.** We evaluate multiple guidance strategies—including actual CFG, linear CFG, Mean-shift guidance, positive CPC guidance, and negative CPC guidance—within the high-noise intervals (the linear regime). For each method, we generate 50,000 images conditioned on the class label “golden retriever” and compute the FD<sub>DINOv2</sub> metric. The results, shown in Figure 37, are consistent with the observations reported in the main text.

**Nonlinear Regime.** We next examine guidance effects in the nonlinear regime using (17) and (18). As shown in Figure 38, guiding with CPCs produces visual effects similar to those of actual CFG—enhancing image sharpness and structure—whereas guidance with conditional PCs often leads to oversaturated colors. This highlights the importance of selectively amplifying class-specific features. We note that our heuristic guidance serves as a conceptual approximation and may not always perfectly align with actual CFG behavior. Additional failure cases will be provided in our code release.

## H CFG in Gaussian Mixture Model

Thus far we’ve been focusing on the setting of linear diffusion models, in which the learned score functions are equivalent to those of a Multivariate Gaussian distribution. From a complementary perspective, several works [20, 19, 7] have studied CFG under the Gaussian mixture model data assumption. However, these works assume each Gaussian cluster has isotropic covariance, which is oversimplified for natural image dataset. In this section, we demonstrate that CFG guidance under Gaussian mixture model can be decomposed in a similar way as the case of linear diffusion model.

Consider unconditional data distribution:

$$p_{\text{data}}(\mathbf{x}) = \sum_{i=1}^K \pi_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad (70)$$

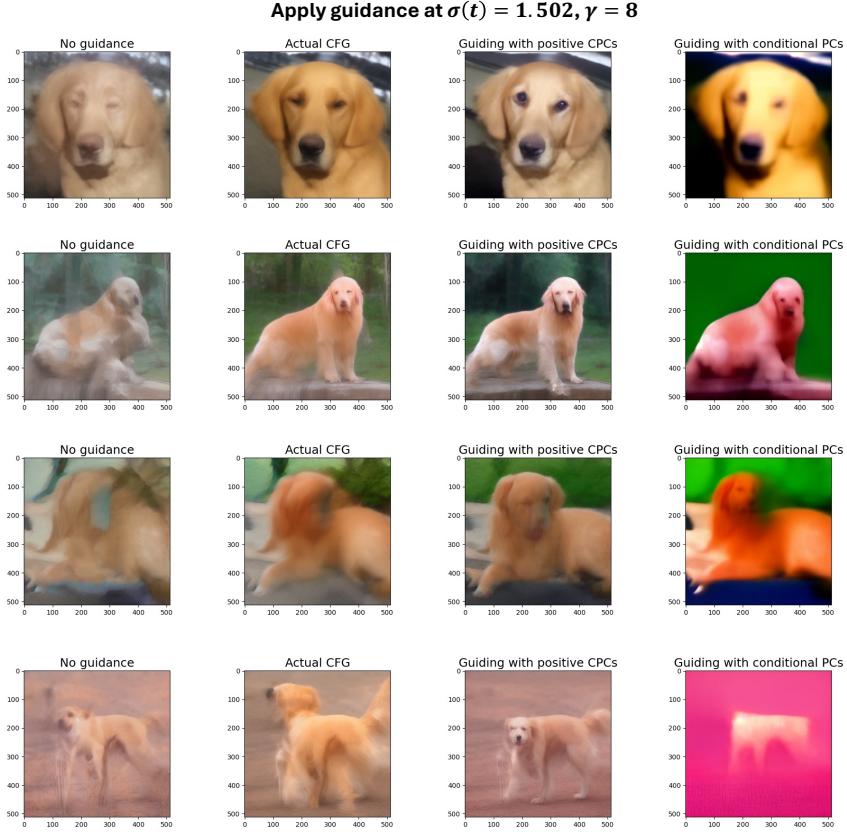


Figure 38: **Effects of CFG in Nonlinear Regime (EDM-2).** Different guidance methods, each with a fixed strength of  $\gamma = 8$ , are applied at  $\sigma(t) = 1.502$ . The samples in each row are generated from the same initial noise.

where  $\boldsymbol{\mu}_i$  and  $\boldsymbol{\Sigma}_i$  are the mean and covariance of the  $i^{th}$  cluster with weight  $\pi_i$ . The noise-mollified data distribution then takes the following form:

$$p(\mathbf{x}; \sigma(t)) = \sum_{i=1}^K \pi_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i + \sigma^2(t) \mathbf{I}). \quad (71)$$

Let  $\boldsymbol{\Sigma}_{\sigma(t),i} := \boldsymbol{\Sigma}_i + \sigma^2(t) \mathbf{I}$ , then the score function of  $p(\mathbf{x}; \sigma(t))$  is:

$$\nabla \log p(\mathbf{x}; \sigma(t)) = \frac{\nabla p(\mathbf{x}; \sigma(t))}{p(\mathbf{x}; \sigma(t))} \quad (72)$$

$$= \frac{\sum_{i=1}^K \pi_i \nabla \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{\sigma(t),i})}{\sum_{i=1}^K \pi_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{\sigma(t),i})} \quad (73)$$

$$= \frac{\sum_{i=1}^K \pi_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{\sigma(t),i}) \boldsymbol{\Sigma}_{\sigma(t),i}^{-1} (\boldsymbol{\mu}_i - \mathbf{x})}{\sum_{i=1}^K \pi_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{\sigma(t),i})} \quad (74)$$

$$= \sum_{i=1}^K w_i(\mathbf{x}) \boldsymbol{\Sigma}_{\sigma(t),i}^{-1} (\boldsymbol{\mu}_i - \mathbf{x}), \quad (75)$$

where  $w_i(\mathbf{x}) = \frac{\pi_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{\sigma(t),i})}{\sum_{i=1}^K \pi_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{\sigma(t),i})}$  representing the posterior probability that  $\mathbf{x}$  belongs to the  $i^{th}$  cluster and  $\sum_{i=1}^K w_i(\mathbf{x}) = 1$ . Let  $\boldsymbol{\Sigma}_i = \mathbf{U}_i \boldsymbol{\Lambda}_i \mathbf{U}_i^T$  be the full SVD where  $\boldsymbol{\Lambda}_i = \text{diag}(\lambda_{i,1}, \dots, \lambda_{i,d})$ ,

by Tweedie's formula, the optimal denoiser of the noise-mollified Gaussian mixture model takes the following form:

$$\mathcal{D}(\mathbf{x}; \sigma(t)) = \mathbf{x} + \sigma^2(t) \nabla \log p(\mathbf{x}; \sigma(t)) \quad (76)$$

$$= \mathbf{x} + \sigma^2(t) \sum_{i=1}^K w_i(\mathbf{x}) \Sigma_{\sigma(t),i}^{-1} (\boldsymbol{\mu}_i - \mathbf{x}) \quad (77)$$

$$= \sum_{i=1}^K w_i(\mathbf{x}) \boldsymbol{\mu}_i + \sum_{i=1}^K w_i(\mathbf{x}) \mathbf{U}_i \tilde{\boldsymbol{\Lambda}}_{\sigma(t),i} \mathbf{U}_i^T (\mathbf{x} - \boldsymbol{\mu}_i), \quad (78)$$

where  $\tilde{\boldsymbol{\Lambda}}_{\sigma(t),i} = \text{diag} \left( \frac{\lambda_{i,1}}{\lambda_{i,1} + \sigma^2(t)}, \dots, \frac{\lambda_{i,d}}{\lambda_{i,d} + \sigma^2(t)} \right)$ . Furthermore, under the Gaussian mixture model assumption, each conditional distribution is a Gaussian distribution and from (7) we know the conditional optimal denoiser of the  $i^{th}$  cluster is:

$$\mathcal{D}(\mathbf{x}; \sigma(t), \mathbf{c}_i) = \boldsymbol{\mu}_i + \mathbf{U}_i \tilde{\boldsymbol{\Lambda}}_{\sigma(t),i} \mathbf{U}_i^T (\mathbf{x} - \boldsymbol{\mu}_i). \quad (79)$$

Without loss of generality, we set the target condition as  $\mathbf{c}_1$ . Then the CFG guidance at timestep  $t$  takes the form:

$$g(\mathbf{x}, t) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x} | \mathbf{c}_1; \sigma(t)) - \nabla_{\mathbf{x}_t} \log p(\mathbf{x}; \sigma(t)) \quad (80)$$

$$= \frac{1}{\sigma^2(t)} (\mathcal{D}(\mathbf{x}_t; \sigma(t), \mathbf{c}_1) - \mathcal{D}(\mathbf{x}_t; \sigma(t))) \quad (81)$$

$$= \frac{1}{\sigma^2(t)} (\mathbf{U}_1 \tilde{\boldsymbol{\Lambda}}_{\sigma(t),1} \mathbf{U}_1^T - \sum_{i=1}^K w_i(\mathbf{x}) \mathbf{U}_i \tilde{\boldsymbol{\Lambda}}_{\sigma(t),i} \mathbf{U}_i^T) (\mathbf{x} - \boldsymbol{\mu}_1) \quad (82)$$

$$+ \frac{1}{\sigma^2(t)} \sum_{i=2}^K w_i(\mathbf{x}) (\mathbf{I} - \mathbf{U}_i \tilde{\boldsymbol{\Lambda}}_{\sigma(t),i} \mathbf{U}_i^T) (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_i). \quad (83)$$

Note that:

- Guidance (82) resembles the CPC guidance  $g_{cpc}(t)$  defined in (12). Different from the linear setting—where the CPC guidance contrasts the posterior covariance of the target class with a single unconditional posterior covariance, here it contrasts the posterior covariance of the target class with a softmax-weighted average of the posterior covariances of all classes.
- Guidance (83) resembles the mean-shift guidance  $g_{mean}(t)$  defined in (12). Different from the linear setting where the mean-shift guidance approximately aligns with  $\boldsymbol{\mu}_c - \boldsymbol{\mu}_{uc}$ , the difference between the conditional and unconditional means, here it instead approximately aligns with a softmax-weighted average of the pairwise differences between the conditional mean (mean of the target class) and the means of every other class.

## I Computing Resources

All experiments are performed on A100 GPUs with 80 GB memory. The diffusion model used in this work typically requires 3 days for sufficient training.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: In the abstract and introduction, we claim we (i) analyze how CFG works in linear setting and (ii) investigate how the insights from linear analysis extend to real-world nonlinear models. The former claim is studied in section 3 while the latter is studied in section 4.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: In section 4.2 we explicitly mention that the analysis in linear regime is only a proxy. The exact mechanism remains unknown due the black-box nature of deep networks.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: We provide the assumption and the theory completely in section [3.2](#).

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We've provided the general information for the experiment setups in both main text and appendix. We promise to release code upon publication to enable reimplementation.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Although we do not provide code in the submission, we promise to release it upon publication of the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have provide detailed information on the datasets we use, models we use and other hyperparameters including number of sampling steps, type of sampler and many others.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Although we do not provide error bar, all the quantitative experiments are performed according to the standard. For example, all the FD scores are calculated over 50,000 randomly generated images.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: We provide the related information in appendix I.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification: All the authors have read the NeurIPS Code of Ethics. We confirm the research conform with the code in every respect.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[NA\]](#)

Justification: Our work is an empirical and theoretical work on understanding the mechanism of CFG. It has more scientific contributions rather than societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our work is an empirical and theoretical work on understanding the mechanism of CFG. It doesn't have a high risk of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use the public released dataset and models, which properly credited the license.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](http://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Our work is an empirical and theoretical work on understanding the CFG mechanism. We don't release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.