LEARNING TO BE GUIDED: SHAPING LATENT GEOMETRY FOR CLASSIFIER-FREE GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Classifier-free guidance (CFG) is a standard component of conditional diffusion models, and guided performance has become the *de facto* measure for state-of-the-art models. In latent diffusion, recent work designs latent spaces that are easier to model. However, under CFG, generation additionally depends on how conditions are arranged in the latent space, revealing a factor that remains largely overlooked. We study this arrangement under discrete class conditions, termed *class geometry*, and investigate what class geometry allows a latent space to work well with CFG. By analyzing the oracle CFG correction and class distinguishability under noise, we show that greater inter-class separation and smaller intra-class spread are favorable for CFG, and derive a noisy posterior objective whose gradient aligns with the oracle correction. Guided by the analysis, we propose Prototype Regularization (PR), a lightweight surrogate for the intractable noisy posterior, which applies a cosine prototype classifier to spatially pooled latents. Under a controlled training budget (80 epochs), adding PR to an LDM baseline reduces guided gFID on ImageNet from 3.31 to 2.28, surpassing a DINOv3-aligned VA-VAE (2.44) with negligible overhead. Through rich extensions, we show that PR generalizes to various label choices, data modalities, and condition signals. Together, our findings highlight class geometry as a critical factor in guided generation and take a first step toward CFG-compatible representation learning.

1 INTRODUCTION

Diffusion models (Ho et al., 2020; Song et al., 2021; Lipman et al., 2023) have become a predominant paradigm in generative modeling, with recent advances largely driven by conditional generation on auxiliary signals such as class labels and text (Rombach et al., 2022; Esser et al., 2024). Classifier-free guidance (CFG) (Ho & Salimans, 2021) is a standard technique for strengthening the conditioning: during sampling, a single model makes conditional and unconditional predictions, and their displacement (i.e., the CFG correction) is amplified by a guidance scale of $\omega > 1$. CFG improves generation quality and has become a crucial component of state-of-the-art models: headline results are generally reported under CFG, usually with a large margin over unguided sampling. Consequently, guided performance has become the *de facto* measure of a conditional diffusion model.

In parallel, latent diffusion models (LDMs) (Rombach et al., 2022) have become the prevailing approach for generating high-dimensional data: an autoencoder first compresses data into a compact latent space, and a diffusion model is then trained in this latent space. A growing line of research designs latent spaces that are easier to model, either by aligning latents with pretrained representations (Yao et al., 2025; Leng et al., 2025) or by adopting pretrained representations directly as the latent space (Zheng et al., 2026; Shi et al., 2026), while a parallel line of analysis studies which properties of a latent space make it easier for diffusion models to learn (Skorokhodov et al., 2025; Zhong et al., 2026; Liu et al., 2026). These efforts largely focus on the learnability of the marginal latent distributions $p_t(z_t)$. Under CFG, however, the guided score for a condition c (e.g., a class label or text embedding) takes the form $\nabla_{z_t} \log p_t(z_t) + \omega \nabla_{z_t} \log p_t(c \mid z_t)$ under exact score estimation, where ω amplifies the condition-specific log-likelihood gradient. With this additional dependence on the condition, accurate modeling of the marginal distributions $p_t(z_t)$ alone does not necessarily translate into better generation performance under guidance.

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

The key lies in the guidance term $\nabla_{z_t} \log p_t(c \mid z_t)$, which pushes z_t toward regions where c is more probable, and its effectiveness depends on how distinguishable latents of different conditions remain after noise corruption. This adds a requirement beyond a well-modeled $p_t(z_t)$: a suitable geometric arrangement of conditions in the latent space. For tractable analysis, we study this problem under the setting of discrete class conditions, and term the arrangement of classes in the latent space *class geometry*. We ask: *what class geometry allows a latent space to work well with CFG?*

We address this question by analyzing CFG correction through the lens of class geometry (Sec. 2). Specifically, we derive a closed form of the oracle CFG correction under a Gaussian class model, which contrasts the target class mean against competing class means, weighted by the noisy class posterior $p_t(c \mid z_t)$. We then show that, at fixed total variance, reallocating intra-class variance to inter-class variance makes classes more distinguishable under noise, so the target-class posterior concentrates earlier in denoising. Finally, we turn this finding into a cross-entropy objective on the noisy class posterior, whose input gradient is proportional to the oracle correction.

Guided by our analysis, we propose Prototype Regularization (PR), which replaces the intractable noisy posterior with a clean-latent surrogate (Sec. 3). Specifically, PR applies a cosine prototype classifier to spatially pooled latents, and we show that its gradient forms a target-versus-competitor contrast analogous to that in the oracle CFG correction. On ImageNet class-conditional generation, adding PR to an LDM baseline reduces guided gFID from 3.31 to 2.28, outperforming DINOv3-aligned (Siméoni et al., 2025) VA-VAE (2.44) (Yao et al., 2025) with negligible overhead (Sec. 4.2).

Moreover, transplanting PR class centers into LDM latents recovers most of the guided gain, suggesting that the improvement stems largely from class geometry (Sec. 4.3). We demonstrate the versatility of PR via extensions to pseudo labels, video latents, and continuous conditions (Sec. 4.4, 4.5, 4.6). In essence, PR encourages a cluster geometry in the latent space based on categorical priors. Our results suggest that PR extends beyond ground-truth labels whenever this geometry stays informative under noise. Together, our findings highlight class geometry as a critical factor in guided generation, and take a first step toward CFG-compatible representation learning.

Our contributions are threefold:

- We approach CFG-compatible representation learning by analyzing CFG through the lens of class geometry. Under a Gaussian class model, we show that greater inter-class separation and smaller intra-class spread keep classes distinguishable under noise, and we derive a noisy posterior objective whose negative gradient aligns with the oracle correction.
- We propose Prototype Regularization (PR), a lightweight surrogate for the noisy posterior objective that shapes class geometry in the latent space, improving guided generation.
- We demonstrate the generalization of PR to pseudo labels, video latents, and continuous conditions, suggesting its effectiveness as a geometry regularizer under categorical priors.

2 From CFG Geometry to Representation Learning

2.1 Preliminaries

Latent Diffusion Models. A latent diffusion model (LDM) (Rombach et al., 2022) operates in a compressed latent space learned by an autoencoder. An encoder $\mathcal{E}_\theta$ maps the raw data $x \in \mathbb{R}^D$ to a compact latent code $z = \mathcal{E}_\theta(x) \in \mathbb{R}^d$ with $d \ll D$, and a decoder $\mathcal{D}_\phi$ reconstructs $\hat{x} = \mathcal{D}_\phi(z)$. For images, the autoencoder is trained with $\mathcal{L}_{\mathrm{LDM}} = \mathcal{L}_{\mathrm{Recon}} + \lambda_{\mathrm{KL}} \mathcal{L}_{\mathrm{KL}} + \lambda_{\mathrm{LPIPS}} \mathcal{L}_{\mathrm{LPIPS}} + \lambda_{\mathrm{GAN}} \mathcal{L}_{\mathrm{GAN}}$. The diffusion model is then trained in the resulting latent space, where the latents are typically standardized via scalar factors or per-channel statistics.

Rectified Flow. Rectified flow (Liu et al., 2023; Lipman et al., 2023) is a widely adopted training objective for diffusion models. For a clean latent $z \in \mathbb{R}^d$, we construct the linear interpolant $z_t = t\, z + (1-t)\, \epsilon$, with independent noise $\epsilon \sim \mathcal{N}(0, I)$, $t \in (0, 1)$ and train a neural network $v_\psi(z_t, t)$ to predict the velocity target $v = z - \epsilon$ by minimizing $\mathcal{L}_{\mathrm{RF}} = \mathbb{E}_{z,\, \epsilon,\, t}\big[\|v_\psi(z_t, t) - (z - \epsilon)\|^2\big]$.

Classifier-Free Guidance. Classifier-free guidance (CFG) (Ho & Salimans, 2021) amplifies the effect of conditioning during sampling to obtain higher-fidelity generations. A single network is trained with condition dropout to produce both a conditional prediction $\widehat{v}_c$ and an unconditional prediction $\widehat{v}_u$. The guided velocity is $\widehat{v}_{\mathrm{cfg}} = \widehat{v}_u + \omega\,(\widehat{v}_c - \widehat{v}_u)$, where $\omega > 1$ is the guidance scale.

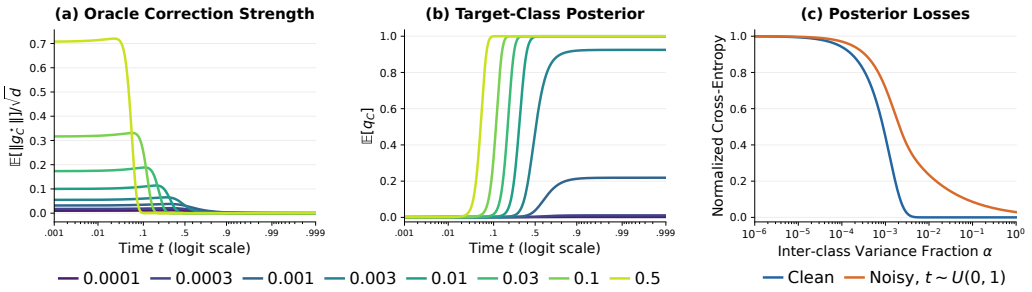


Figure 1: Simulations under the Gaussian class model with $K = 1000$ classes and $d = 8192$ dimensions (Eq. 4). (a) Oracle CFG correction strength along time t ($t = 1$ is clean), for varying inter-class variance fraction α. The correction vanishes once the target-class posterior concentrates. (b) Expected target-class posterior under forward corruption along t. Larger α concentrates the posterior earlier. (c) Posterior losses (normalized by $\log K$) versus α, on clean-only and uniform-time corrupted latents. Larger α yields smaller entropy (i.e., more confident posteriors).

2.2 A Geometric View of Classifier-Free Guidance

How Class Geometry Enters CFG. Under rectified flow (Liu et al., 2023), let $v_c^\star = \mathbb{E}[v \mid z_t, t, C = c]$ and $v_u^\star = \mathbb{E}[v \mid z_t, t]$ be the Bayes-optimal conditional and unconditional predictors. Substituting $v = (z - z_t)/(1 - t)$, the oracle CFG correction $g_c^\star(z_t, t)$ reduces to:

$$g_c^\star(z_t, t) := v_c^\star - v_u^\star = \frac{\mathbb{E}[z \mid z_t, t, C = c] - \mathbb{E}[z \mid z_t, t]}{1 - t}. \tag{1}$$

Overall, the correction captures how conditioning on the target class changes the posterior estimate of the clean latent. To gain concrete intuition on how class geometry shapes $g_c^\star$, we derive a closed form of Eq. 1 under a simplified setting. Suppose the condition $c \in \{1, \dots, K\}$ indexes K discrete classes, each with a Gaussian class-conditional prior $p(z \mid c) = \mathcal{N}(z;\, \mu_c,\, \sigma^2 I)$ with shared within-class variance σ^2, which we refer to as the *Gaussian class model*. Let $q_j = \Pr(C = j \mid z_t, t)$ denote the posterior class probability. We obtain:

$$g_c^\star(z_t, t) = \frac{1 - t}{(1-t)^2 + t^2\sigma^2} \sum_{j \neq c} q_j \, (\mu_c - \mu_j). \tag{2}$$

Therefore, the oracle CFG correction can be expressed as a posterior-weighted sum of displacements from the competing class means toward the target mean, with σ^2 shaping both the prefactor and the posterior weights. The center displacements $\mu_c - \mu_j$ define the class contrasts, while the posterior weights q_j determine which contrasts are emphasized at each noisy state z_t. When the target class dominates the posterior (i.e., $q_c \to 1$), the weights of all competing classes vanish, eliminating the target-versus-competitor contrast. This behavior is illustrated in Fig. 1(a) and Fig. 1(b).

Class Distinguishability under Noise. To understand how class geometry governs this posterior concentration, we examine class distinguishability under noise corruption. Under the same Gaussian model, the class-conditional distribution of z_t is $p(z_t \mid c) = \mathcal{N}\big(z_t; t\mu_c, (t^2\sigma^2 + (1-t)^2)I\big)$, so the squared standardized distance between classes c and j is given by:

$$\delta_{cj}^2(t) = \frac{t^2 \|\mu_c - \mu_j\|^2}{t^2\sigma^2 + (1-t)^2}. \tag{3}$$

To isolate the effect of class geometry from changes in latent scale, we consider a controlled family of latent distributions. With class priors held fixed, we choose the reference class means satisfying $\mathbb{E}_C[\mu_C^{\text{ref}}] = 0$ and $\mathbb{E}_C\|\mu_C^{\text{ref}}\|^2 = d$, where d is the latent dimension, and set:

$$\mu_c(\alpha) = \sqrt{\alpha}\, \mu_c^{\text{ref}}, \qquad \sigma^2(\alpha) = 1 - \alpha, \qquad \alpha \in (0, 1). \tag{4}$$

Under this controlled family, the total variance remains d, with inter-class and intra-class contributions of αd and $(1 - \alpha)d$, respectively. Increasing the inter-class variance fraction α separates class centers while reducing intra-class variance, thereby increasing $\delta_{cj}^2(t)$ at every fixed $t \in (0, 1)$. Consequently, classes become distinguishable under stronger corruption, and the target-class posterior concentrates earlier along the denoising process, as illustrated in Fig. 1(b). This identifies variance allocation, rather than latent scale alone, as a way to shape the noisy class posterior.

2.3 A CFG-MOTIVATED REPRESENTATION OBJECTIVE

Motivated by how CFG contrasts classes through the noisy posterior probabilities in Eq. 2, we seek representations that make the target class more distinguishable from its competitors under corruption. We express this geometric preference through an ideal posterior cross-entropy objective:

$$\mathcal{L}_{\text{post}}(t) = \mathbb{E}_{z,c,\epsilon}\left[-\log q_c\big(tz + (1-t)\epsilon, t\big)\right] = H(C \mid Z_t), \tag{5}$$

where q_c is the true noisy class posterior, and $H(C \mid Z_t)$ denotes the conditional entropy of the class label given the corrupted latent. This objective is directly related to the oracle correction $g_c^\star$ via its input gradient. With the posterior function held fixed, the per-sample descent direction satisfies:

$$\nabla_z \log q_c\big(tz + (1-t)\epsilon, t\big) = t\nabla_{z_t} \log q_c(z_t, t) = \frac{t^2}{1-t}\, g_c^\star(z_t, t), \quad t \in (0, 1). \tag{6}$$

Thus, the ideal posterior objective provides a per-sample latent descent direction proportional to the oracle CFG correction. In addition, the objective also matches the geometric preference identified in the fixed-variance family (Eq. 4): increasing α raises the effective class signal-to-noise ratio, making $\mathcal{L}_{\text{post}}(t)$ non-increasing at each noise level (Appendix E.8). Fig. 1(c) compares the posterior loss on clean latents ($t = 1$) with its average over corruption levels $t \sim \mathcal{U}(0, 1)$. Both losses are divided by $\log K$ (i.e., the entropy of a uniform distribution over the K classes), so a value of one corresponds to complete class uncertainty in this equal-prior setting. As the inter-class variance fraction α increases, both losses decrease, indicating that class identity becomes more distinguishable. Meanwhile, the clean posterior loss saturates earlier than its noise-averaged counterpart.

Different choices of the corruption-time distribution define a family of posterior objectives, including the uniform-time and clean-only variants above. In practice, uniform-time supervision can overregularize the representation, whereas clean-only supervision already yields substantial gains in guided generation (Appendix D.1). The connection between posterior objectives and CFG correction provides a basis for designing practical regularizers for CFG-compatible representations.

3 PROTOTYPE REGULARIZATION

Sec. 2.3 motivates $\mathcal{L}_{\text{post}}$ as an ideal objective for CFG-compatible representations, but it cannot be optimized directly: the true posterior q_c is defined by the very latent distribution being learned. Since the diffusion model is trained in the latent space fixed by the autoencoder, we instead optimize a practical surrogate during autoencoder training, termed *Prototype Regularization* (PR). PR makes two simplifications: it adopts clean-only supervision ($t = 1$), and it replaces q_c with a lightweight prototype classifier that mirrors the posterior of a Gaussian class model.

Prototype classifier on pooled latents. Given a latent $z \in \mathbb{R}^{d_c \times h \times w}$, we obtain a global feature $\bar{z} = \text{AvgPool}(z) \in \mathbb{R}^{d_c}$ and associate each class j with a learnable prototype $p_j \in \mathbb{R}^{d_c}$, a proxy for its class mean. With $\tilde{z} = \bar{z}/\|\bar{z}\|$ and $\tilde{p}_j = p_j/\|p_j\|$, the prototype posterior and the PR loss are:

$$\pi_j(z) = \frac{\exp(\tilde{z}^\top \tilde{p}_j/\tau)}{\sum_{k=1}^K \exp(\tilde{z}^\top \tilde{p}_k/\tau)}, \qquad \mathcal{L}_{\text{PR}} = -\log \pi_c(z), \tag{7}$$

where c is the class label and $\tau > 0$ is a fixed temperature. Average pooling restricts PR to the spatial mean of z, leaving the residual $z - \bar{z}$ unconstrained; this avoids pulling the full latent toward a class template, which may constrain layout and reduce diversity. Normalization makes the objective depend only on angular alignment, which excludes the trivial minimizer of inflating feature norms.

Connection to CFG correction. Under the Gaussian class model (Sec. 2.2) with equal priors and equal-norm class means, the clean posterior reduces to $q_j \propto \exp(\mu_j^\top z/\sigma^2)$. PR adopts the same form on normalized features, with learnable prototypes in place of μ_j. Consequently, its negative gradient inherits the target-versus-competitor contrast:

$$-\nabla_{\tilde{z}} \mathcal{L}_{\text{PR}} = \frac{1}{\tau}\Big(\tilde{p}_c - \sum_{j=1}^K \pi_j \tilde{p}_j\Big) = \frac{1}{\tau}\sum_{j \neq c} \pi_j\, (\tilde{p}_c - \tilde{p}_j), \tag{8}$$

which mirrors Eq. 2 term by term, with prototypes in place of class means and π_j in place of q_j: each sample is pulled toward its target and away from its competitors, with more weight on those most easily confused with it. We compare PR variants, including noisy supervision, in Appendix D.

Training objective. We train the autoencoder with:

$$\mathcal{L} = \mathcal{L}_{\mathrm{LDM}} + \lambda_{\mathrm{PR}}\mathcal{L}_{\mathrm{PR}}, \tag{9}$$

where the reconstructive terms in $\mathcal{L}_{\mathrm{LDM}}$ retain instance-specific information alongside the class structure from PR. Following Esser et al. (2021) and Rombach et al. (2022), λ_{PR} is set adaptively by balancing encoder last-layer gradient norms against the reconstruction term.

Beyond class labels. PR only requires a partition of the data, so ground-truth labels can be replaced by pseudo-labels from clustering pretrained features, with one prototype per cluster. Moreover, the regularization labels need not match the generation condition: once the autoencoder is trained with PR, the diffusion model can be trained on other downstream conditions, discrete or continuous. We show that PR remains effective in these settings (Sec. 4.4, 4.6).

4 Experiment

4.1 Experimental Setup

Our primary testbed is class-conditional image generation on ImageNet (Deng et al., 2009) at 256×256 resolution. We build on the VA-VAE codebase (Yao et al., 2025) and use a convolution-based encoder/decoder with a latent shape of $32 \times 16 \times 16$, paired with LightningDiT-XL/1 as the stage-2 backbone. We use the same number of training epochs for all methods (20 epochs for stage-1 and 80 epochs for stage-2) for fair comparison. All ImageNet experiments in the main text are trained on $8\times$ NVIDIA H100 GPUs; a subset of appendix experiments uses 2 GPUs (with 10 epochs for stage-1 and 30 epochs for stage-2). Full hyperparameter details are provided in Appendix C.

We use generation FID (gFID) (Heusel et al., 2017) as our primary metric, which measures how closely generated images match the training distribution. For all results with classifier-free guidance, we report gFID at the best guidance scale $\omega^\star$, selected via a guidance scale sweep using gFID computed on 5K samples. When comparing against baselines, we follow the standard practice and additionally report sFID, Inception Score (IS), Precision, and Recall using the ADM evaluation toolkit (Dhariwal & Nichol, 2021). Most reported metrics are computed on 50K samples. Setups for Sec. 4.5 and Sec. 4.6 are described in their respective subsections.

4.2 Class-Conditional Image Generation with Ground-truth Labels

We compare PR against LDM (Rombach et al., 2022) and VA-VAE (Yao et al., 2025). We trained two VA-VAE variants, aligned with DINOv3-B and DINOv3-L (Siméoni et al., 2025), respectively. Our method integrates PR into LDM training. On reconstruction fidelity, we report PSNR, LPIPS, and rFID. Overall, all methods perform comparably, with VA-VAE holding a marginal edge. PR therefore does not sacrifice reconstruction quality for class geometry. Meanwhile, PR raises Purity@10 (i.e., the fraction of each latent's 10 nearest neighbors that share its class) from 0.020 (LDM) to 0.279, about $4\times$ that of the strongest VA-VAE variant (0.071), indicating substantially more class-consistent local neighborhoods. PR achieves this at negligible cost, whereas VA-VAE slows stage-1 training throughput by about 25% due to the additional representation alignment.

On the generation side, the gains of PR concentrate on guided settings. Without CFG, PR improves over LDM ($13.09 \rightarrow 11.59$) but trails both VA-VAE variants. With CFG, the ranking reverses: PR achieves the best gFID (2.28), IS (269.85), and Recall (0.585) among all methods. Moreover, the gFID ranking with CFG is consistent with the Purity ranking, while the ranking without CFG does not. This suggests that class geometry primarily governs how much a model benefits from guidance rather than its unguided quality, which we investigate in Section 4.3. PR shows slightly lower Precision than LDM (0.808 vs. 0.859), consistent with its higher Recall.

Figure 2 examines the latent spaces from three complementary angles: (a) We decompose the energy of per-channel normalized latents into a global mean, inter-class variation, and intra-class variation. In LDM, over 90% of the energy belongs to the intra-class portion. VA-VAE mildly increases inter-class energy, but shifts a large portion to the global mean, a spatial template shared by all latents. In contrast, PR reallocates intra-class energy to inter-class variance, yielding the largest inter-class share among all methods. (b) We measure the average posterior $\mathbb{E}[q_c(z_t)]$ of the target class on both forward-noised validation latents and intermediate states along CFG sampling trajectory, where

Table 1: Results of class-conditional generation on ImageNet. All quality metrics are evaluated with 50K samples. Purity measures the fraction of each latent's K = 10 nearest neighbors that share its class, averaged over 50K normalized validation latents. Best results are highlighted in **bold**.

Method	Stage 1					Stage 2					
	Reconstruction			Geometry	Speed	w/o CFG	w/ CFG				
	rFID↓	PSNR↑	LPIPS↓	Purity↑	Steps/s↑	gFID↓	gFID↓	sFID↓	IS↑	Precision↑	Recall↑
LDM	0.308	27.07	0.101	0.020	**2.16**	13.09	3.31	**4.68**	237.73	**0.859**	0.504
VA-VAE (DINOv3-B)	0.280	**27.09**	**0.099**	0.053	1.65	9.76	2.64	4.94	236.02	0.828	0.557
VA-VAE (DINOv3-L)	**0.275**	27.06	0.099	0.071	1.61	**9.01**	2.44	4.83	242.17	0.816	0.564
LDM + PR (Ours)	0.277	26.95	0.101	**0.279**	2.13	11.59	**2.28**	4.87	**269.85**	0.808	**0.585**

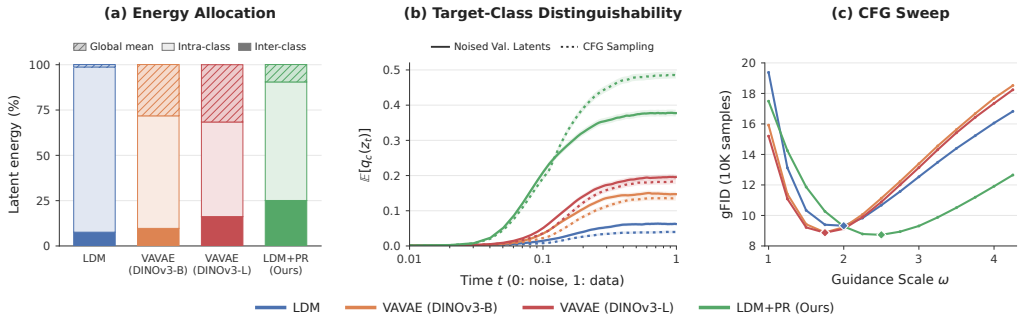


Figure 2: (a) Energy decomposition of per-channel normalized latents into global mean $\|\mu\|^2$, inter-class variance $\mathbb{E}_c\|\mu_c - \mu\|^2$, and intra-class variance $\mathbb{E}\|z - \mu_c\|^2$. (b) Average target-class posterior $\mathbb{E}[q_c(z_t)]$ along the diffusion trajectory, with z_t obtained by noised validation latents (solid) or taken from intermediate states of CFG sampling (dotted). (c) gFID (10K samples) versus guidance scale ω; diamonds mark the best scale $\omega^\star$. PR allocates more energy to inter-class variance, makes the target class identifiable earlier and more confidently, and tolerates stronger guidance.

$q_c(z_t)$ denotes the posterior probability of class c at z_t. PR yields a substantially higher posterior throughout, so the target class becomes identifiable under noise. CFG thus steers samples toward class-consistent regions more effectively. (c) Sweeping the guidance scale, PR reaches its optimum at a larger $\omega^\star$ and degrades more slowly beyond it, resulting in a wider near-optimal range. This suggests that class geometry shaped via PR allows the model to benefit from stronger guidance.

4.3 Understanding the Gains through Center Transplantation

Figure 2(a) shows that PR primarily reallocates its energy to inter-class variance (i.e., to the class centers). This raises a natural question: *can an unregularized latent space benefit from the class geometry of PR in a plug-and-play manner?* To answer this, we transplant PR class centers into the LDM latent space. For a latent z of class c, we replace its native center with a target center, $z' = z - \mu_c^{\mathrm{LDM}} + \tilde{\mu}_c$, which preserves intra-class deviations, and train a stage-2 model on z'. We do the inverse operations before decoding to ensure representation consistency. We transplant PR centers in two shapes: the full center and its spatially pooled version. For fair comparison, we include two controls: LDM with its own centers rescaled to match the inter-class energy of the full PR centers, and transplanted random directions with PR's per-class norms. As shown in Tab. 2, transplanting PR centers improves guided generation (gFID $3.31 \rightarrow 2.52$ with pooled centers), while the unguided setting changes only marginally ($13.09 \rightarrow 12.65$). Increasing the inter-class energy of LDM alone yields a partial gain (2.89), and PR centers improve further beyond this energy-matched control. In contrast, random directions provide little benefit with CFG and even hurt unguided generation. These results indicate that PR centers have a better arrangement, and the gain does not stem from increased inter-class energy alone. Besides, pooled centers outperform spatially resolved ones (2.52 vs. 2.63), so the useful class signal is mainly a global channel-wise offset.

Interestingly, all models trained on transplanted spaces reach a larger $\omega^\star$, indicating that a latent space with separated class centers can take stronger guidance. Finally, center transplantation recovers most of PR's guided gain but only a small portion of its unguided gain, and thus does not fully reproduce the benefits of PR. This is likely because transplantation only alters inter-class separation and leaves intra-class spread untouched, whereas PR reshapes both during stage-1 training.

Table 2: Center transplantation on LDM latents ($z' = z - \mu_c^{\mathrm{LDM}} + \tilde{\mu}_c$). Rescaled LDM centers match the inter-class energy of full PR centers. Random centers match the pooled PR energy. Spatially constant centers are broadcast over the spatial grid. $\omega^\star$ denotes the best CFG scale.

Base Latent	Center Construction	Center Shape	gFID@50K↓		$\omega^\star$
			w/o CFG	w/ CFG	
LDM + PR	Native	–	11.59	2.28	2.5
LDM	Native	–	13.09	3.31	2.0
LDM	LDM (Rescaled)	$32 \times 16 \times 16$	12.58	2.89	2.0
LDM	PR	$32 \times 16 \times 16$	12.79	2.63	3.5
LDM	PR	$32 \times 1 \times 1$	12.65	2.52	4.0
LDM	Random Directions + PR Norms	$32 \times 1 \times 1$	13.74	3.16	4.5

4.4 Generalization under Cluster Labels and Non-identical Label Sets

Building on PR's performance gains on ground-truth class labels, we further investigate whether PR remains effective on other types of labels, and whether the two stages must share the same label set. To this end, we construct unsupervised labels by clustering the `[CLS]` tokens of DINOv3-L (Siméoni et al., 2025) with spherical K-means, setting $K{=}1000$ to match the number of ImageNet classes. We then evaluate all combinations of stage-1 PR labels and stage-2 conditioning labels (Tab. 3). When conditioning on DINO clusters, we sample cluster labels in proportion to cluster sizes during evaluation.

Table 3: PR with different label set pairings for stage-1 regularization and stage-2 conditioning. We report guided gFID / unguided gFID below. The best result per-column is in **bold**.

PR Labels	Stage 2 Condition	
	Ground-truth	DINO Clusters
None	3.31 / 13.09	3.58 / 13.31
Ground-truth	**2.28** / 11.59	**2.42** / 11.82
DINO Clusters	2.34 / **11.44**	2.55 / **11.47**

Using ground-truth labels in both stages yields the best guided generation (2.28). Replacing the stage-1 labels with DINO clusters recovers most of this gain (2.34, about 94% of the improvement over the no-PR baseline) and even improves unguided generation (11.44 vs. 11.59). When stage-2 conditions on DINO clusters, PR also substantially improves over the no-PR baseline ($3.58 \rightarrow 2.42$ and 2.55 with CFG). Notably, under DINO conditioning, PR regularized with ground-truth labels performs best with CFG (2.42), even though its stage-1 labels differ from the stage-2 condition. These results show that PR neither relies on ground-truth labels nor requires the two stages to share the same label set, making it applicable to settings without human annotations.

4.5 Generalization to Video Latents

We further examine whether PR generalizes to video latents. We use UCF-101 (Soomro et al., 2012), a human action video dataset with 101 action classes, and adopt LeanVAE-16 (Cheng & Yuan, 2025) as the base autoencoder. This VAE is trained with reconstruction objectives and compresses each 17-frame 256×256 clip into a latent of shape $5 \times 16 \times 32 \times 32$ ($T \times C \times H \times W$). Starting from the official VAE checkpoint, we fine-tune for 3 epochs using either reconstruction alone or reconstruction with PR; for video latents, PR's average pooling spans both the spatial and temporal dimensions. We then fine-tune a Latte-XL/2 model (Ma et al., 2025) on each of three latent spaces: the official VAE, the VAE fine-tuned with reconstruction alone, and the VAE fine-tuned with PR. All three Stage 2 runs are initialized from the EMA weights of the same Latte-XL/2 checkpoint provided by LeanVAE and fine-tuned for 50K steps with a global batch size of 10. We also evaluate the VAE and Latte-XL/2 checkpoints provided by LeanVAE without further training as a reference. We report FVD (over 2048 clips), which measures the distance between the generated and real video distributions, and action classification accuracy from a pre-trained UCF-101 classifier, which measures the class consistency of conditional generation. As shown in Tab. 4, PR improves FVD in both settings with on-par accuracy, reducing FVD from 152.7 to 140.1 with CFG and from 287.3 to 275.7 without CFG. These results suggest that the benefit of PR is not specific to images and extends to latent spaces of other modalities.

Table 4: Generalization to video latents on UCF-101 (Soomro et al., 2012) using LeanVAE-16 (Cheng & Yuan, 2025) and Latte-XL/2 (Ma et al., 2025). *Official* denotes the VAE and Latte-XL/2 checkpoints provided by LeanVAE, used without further training. We report FVD and action classification accuracy with and without CFG. The best results are in **bold**.

Stage 1 (LeanVAE)	Stage 2 (Latte-XL/2)	w/ CFG		w/o CFG	
		FVD↓	Accuracy (%)↑	FVD↓	Accuracy (%)↑
Official	Official	167.0	92.9	340.2	87.2
Official	Fine-Tuned	161.4	98.5	302.0	92.7
Fine-Tuned (Recon Only)	Fine-Tuned	152.7	**99.0**	287.3	**93.0**
Fine-Tuned (PR, Ours)	Fine-Tuned	**140.1**	98.9	**275.7**	**93.0**

Table 5: ImageNet text-to-image generation using different Stage-1 latent representations. All metrics use 50K images generated from held-out validation captions. All alignment to CLIP-L adopts its text embeddings. CLIP denotes mean cosine similarity of CLIP embeddings calculated from text captions and generated images. All cluster labels use $K = 1000$. The best results are in **bold**.

Method	w/ CFG			w/o CFG		
	gFID Train↓	gFID Val↓	CLIP↑	gFID Train↓	gFID Val↓	CLIP↑
LDM	3.15	2.90	0.2775	10.38	7.52	0.2577
VA-VAE (DINOv3-L)	2.53	2.67	0.2814	**8.02**	**5.65**	**0.2620**
Cosine Alignment (CLIP-L)	2.93	3.04	0.2799	9.52	6.84	0.2590
Cosine Alignment (DINOv3-L)	2.86	2.78	0.2784	9.35	6.74	0.2585
LDM + PR (GT Labels)	**2.51**	**2.35**	0.2811	9.63	6.87	0.2570
LDM + PR (CLIP-L Clusters)	2.52	2.42	0.2813	10.15	7.36	0.2578
LDM + PR (DINOv3-L Clusters)	2.53	2.46	**0.2818**	9.41	6.72	0.2575

4.6 Generalization to Continuous Conditioning

Finally, we examine whether PR benefits generation under continuous conditioning signals. We consider text-to-image generation on ImageNet: we caption each image with Florence-2-large (Xiao et al., 2024) and encode the caption into a CLIP ViT-L/14 text embedding (Radford et al., 2021), which replaces the class embedding in LightningDiT (Yao et al., 2025) through a small MLP adapter. Besides LDM and VA-VAE (DINOv3-L), we compare against two continuous alignment baselines, which align latents via cosine similarity to pre-trained features projected to the channel dimension by PCA. We apply 1000-class PR with ground-truth classes, clusters of CLIP-L text embeddings, and DINOv3-L `[CLS]` tokens. All metrics are computed on 50K images generated from held-out validation captions. As shown in Table 5, PR variants achieve the best results on all guided metrics. Notably, with CFG, clustering a representation consistently outperforms aligning to it continuously (2.42 vs. 3.04 for CLIP-L and 2.46 vs. 2.78 for DINOv3-L). Without CFG, VA-VAE performs best, mirroring the class-conditional results in Tab. 1. We attribute the guided gains of PR to two factors: ImageNet images are organized around discrete object categories even when described by free-form captions, so a categorical prior captures the dominant structure of the condition; and a distinct categorical geometry keeps the condition identifiable under noise (Figure 2b), which makes guidance easier for the diffusion model to exploit. Overall, PR extends to continuous conditioning as long as its categorical geometry remains informative under noise.

5 Related Work

Latent Representation for Diffusion Models. Latent diffusion models (LDMs) adopt a two-stage paradigm (Rombach et al., 2022): an autoencoder maps high-dimensional data to a compact latent space, and a diffusion model learns to generate within it. Subsequent work expands the design space by widening the latent channels (Esser et al., 2024; Labs et al., 2025) and raising the compression ratio (Chen et al., 2025c). Recent work largely focuses on optimizing latent representations for easier

diffusion modeling. One line of work distills features from pretrained vision foundation models: VA-VAE (Yao et al., 2025) applies such alignment during autoencoder training, and REPA-E (Leng et al., 2025) extends this to end-to-end tuning through REPA alignment (Yu et al., 2024). Similarly, AlignTok (Chen et al., 2025a), RAE (Zheng et al., 2026; Singh et al., 2026), and SVG (Shi et al., 2026) replace the learned encoder with pretrained features and diffuse directly in its feature space. Another line builds semantic structure without any external model: MAETok (Chen et al., 2025b) uses a masked objective in which auxiliary decoders reconstruct masked tokens from visible ones; VTP (Yao et al., 2026) jointly optimizes contrastive, self-supervised, and reconstruction losses.

Diffusability of Latent Representations. A growing literature investigates what properties of the latent space determine its *diffusability* (i.e., how easily the latent distribution can be modeled by diffusion objectives). One line of work attributes diffusability to spectral properties: Skorokhodov et al. (2025) use a scale-equivariance regularizer to reduce the excessive high-frequency content; EQ-VAE (Kouzelis et al., 2025) regularizes the latent through the consistency after geometric transformations; SSVAE (Liu et al., 2026) extends the spectral analysis to video. Others target the latent distribution directly: UL (Heek et al., 2026) replaces the KL penalty with a diffusion-prior regularizer; DMVAE (Ye et al., 2025) aligns the latent distribution to a chosen reference via score matching. More recent work offers multi-factor explanations: PAE (Yue et al., 2026) highlights spatial structure, local manifold continuity, and global semantics; Zhong et al. (2026) identify velocity irreducible variance, semantic separability, and spatial structure as the most stable predictors through a large-scale study.

Classifier-Free Guidance. Classifier-free guidance (CFG) (Ho & Salimans, 2021) improves conditional generation by extrapolating the prediction displacement at inference, amplifying class-specific signal at the cost of diversity and saturation at high guidance scales. A large body of work refines this trade-off on the sampling side (Karras et al., 2024; Kynkäänniemi et al., 2024; Chung et al., 2025; Fan et al., 2025). On the analytical side, Li et al. (2026) study CFG under a linear-Gaussian denoising model, decomposing the guidance vector into a class-mean shift and covariance-dependent contrastive components; Zhao & Schwing (2026) show that CFG implicitly pushes trajectories away from decision boundaries. A few studies touch the representation side: ProtoDiffusion (Baykal et al., 2024) uses pretrained class prototypes as conditioning embeddings in diffusion training; CCUA (Chen et al., 2026) aligns conditional and unconditional noise predictions for long-tailed diffusion.

6 Discussion

Unguided performance. PR primarily targets guided generation, where its gains concentrate; without guidance, it still improves over LDM ($13.09 \rightarrow 11.59$) but trails VA-VAE (Tab. 1). The remaining gap may stem from pooling, which constrains only the channel dimension and leaves spatial structure unregularized, while suppressing high-frequency spatial components has been shown to be important for diffusability (Skorokhodov et al., 2025; Liu et al., 2026). Appendix D.5 shows that applying PR with VA-VAE gives performance gain in both guided and unguided generation.

Design space of PR. PR instantiates the posterior objective through two simplifications, each leaving room for exploration. Clean-latent supervision avoids the over-regularization of noisy supervision (Appendix D.1), though other time distributions may work better. Likewise, the pooled prototype classifier is one of many possible ways to realize the target-versus-competitor contrast.

Scope of analysis. Our derivations are mainly based on the Gaussian class model, which real latents may not satisfy; nevertheless, the empirical posterior trend aligns with the predictions (Fig. 2(b), Tab. 1). Appendix F further relates class geometry to other quantities in CFG.

7 Conclusion

In this work, we study CFG-compatible representation learning through the lens of class geometry. Our analysis links the oracle CFG correction to how distinguishable classes remain under noise, which motivates Prototype Regularization (PR), a lightweight regularizer that shapes class geometry and improves guided generation. PR generalizes across label choices, data modalities, and condition signals, suggesting that aligning latent geometry through categorical priors is a simple yet effective way to design latent spaces for guided generation. Given the wide use of guidance, we believe CFG-compatible representation learning is a promising and underexplored direction.

AI USE STATEMENT

In this work, we used generative AI tools for: generating synthetic data set (Gaussian mixture), assisting in the writing of proofs, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting dataset, supporting qualitative and thematic data analysis, and interpreting results.

We have not used generative AI tools for: developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses.

Additionally, we used generative AI tools for: creating figures, polishing the paper, summarizing existing literature, brainstorming, identifying relevant literature.

We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, private data, or new datasets, and we are not aware of any conflicts of interest. Beyond the ethical concerns widely recognized for generative models in general, such as the potential generation of harmful or illegal content and copyright issues related to training data and generated outputs, we do not foresee additional ethical risks specific to our work.

REPRODUCIBILITY STATEMENT

We have made several efforts to ensure the reproducibility of this work. Hyperparameter settings and other experimental details are given in Appendix C. Full mathematical derivations, together with their underlying assumptions, are provided in Appendix E and F. All experiments are conducted on publicly available datasets and openly released pretrained models, and no closed-source data or models are used. The full code repository will be publicly released upon publication.

REFERENCES

Gulcin Baykal, Halil Faruk Karagoz, Taha Binhuraib, and Gozde Unal. Protodiffusion: Classifier-free diffusion guidance with prototype learning. In *Asian Conference on Machine Learning*, pp. 106–120. PMLR, 2024.

Bowei Chen, Sai Bi, Hao Tan, He Zhang, Tianyuan Zhang, Zhengqi Li, Yuanjun Xiong, Jianming Zhang, and Kai Zhang. Aligntok: Aligning visual foundation encoders to tokenizers for diffusion models. *arXiv preprint arXiv:2509.25162*, 2025a.

Fang Chen, Alex Villa, Gongbo Liang, Fuxing Li, Xiaoyi Lu, and Meng Tang. Contrastive conditional-unconditional alignment for long-tailed diffusion model, 2026. URL https://arxiv.org/abs/2507.09052.

Hao Chen, Yujin Han, Fangyi Chen, Xiang Li, Yidong Wang, Jindong Wang, Ze Wang, Zicheng Liu, Difan Zou, and Bhiksha Raj. Masked autoencoders are effective tokenizers for diffusion models. In *Forty-second International Conference on Machine Learning*, 2025b.

Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 96539–96560, 2025c.

Yu Cheng and Fajie Yuan. Leanvae: An ultra-efficient reconstruction vae for video diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15692–15702. IEEE, 2025.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. CFG++: Manifold-constrained classifier free guidance for diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=E77uvbOTtp.

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*, pp. 248–255. Ieee, 2009.

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

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Weichen Fan, Amber Yijia Zheng, Raymond A. Yeh, and Ziwei Liu. Cfg-zero*: Improved classifier-free guidance for flow matching models. *CoRR*, abs/2503.18886, March 2025. URL https://doi.org/10.48550/arXiv.2503.18886.

Jonathan Heek, Emiel Hoogeboom, Thomas Mensink, and Tim Salimans. Unified latents (ul): How to train your latents. *arXiv preprint arXiv:2602.17270*, 2026.

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.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. EQ-VAE: Equivariance regularized latent space for improved generative image modeling. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=UWhW5YYLo6.

Tuomas Kynkäänniemi, 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. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=nAIhvNy15T.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18262–18272. IEEE, 2025.

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

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Shizhan Liu, Xinran Deng, Zhuoyi Yang, Jiayan Teng, Xiaotao Gu, and Jie Tang. Delving into latent spectral biasing of video vaes for superior diffusability, 2026. URL https://arxiv.org/abs/2512.05394.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *Transactions on Machine Learning Research*, 2025.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. In *International Conference on Learning Representations*, volume 2026, pp. 154506–154537, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=2hEDcA7xy4.

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. URL https://openreview.net/forum?id=PxTIG12RRHS.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4818–4829. IEEE, 2024.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712. IEEE, 2025.

Jingfeng Yao, Yuda Song, Yucong Zhou, and Xinggang Wang. Towards scalable pre-training of visual tokenizers for generation. In *European Conference on Computer Vision*, pp. 661–678. Springer, 2026.

Sen Ye, Jianning Pei, Mengde Xu, Shuyang Gu, Chunyu Wang, Liwei Wang, and Han Hu. Distribution matching variational autoencoder. *arXiv preprint arXiv:2512.07778*, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Zhengrong Yue, Taihang Hu, Mengting Chen, Haiyu Zhang, Zihao Pan, Tao Liu, Zikang Wang, Jinsong Lan, Xiaoyong Zhu, Bo Zheng, et al. What matters for diffusion-friendly latent manifold? prior-aligned autoencoders for latent diffusion. *arXiv preprint arXiv:2605.07915*, 2026.

Xiaoming Zhao and Alex Schwing. Studying classifier (-free) guidance from a classifier-centric perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13271–13279, 2026.

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

Tianxiong Zhong, Xingye Tian, Xuebo Wang, Xin Tao, and Pengfei Wan. Diffusing in the right space: A systematic study of latent diffusability. *arXiv preprint arXiv:2606.03578*, 2026.

APPENDIX OUTLINE

The appendix is organized as follows:

- **Appendix A** presents UMAP visualizations of the learned VAE latents.
- **Appendix B** provides additional visualizations of noisy posterior and oracle velocity field.
- **Appendix C** gives a full details of experiment configurations.
- **Appendix D** reports additional experimental results.
- **Appendix E** derives the equations and claims stated in the main text.
- **Appendix F** derives complementary perspectives on classifier-free guidance (CFG).

A LATENT VISUALIZATIONS

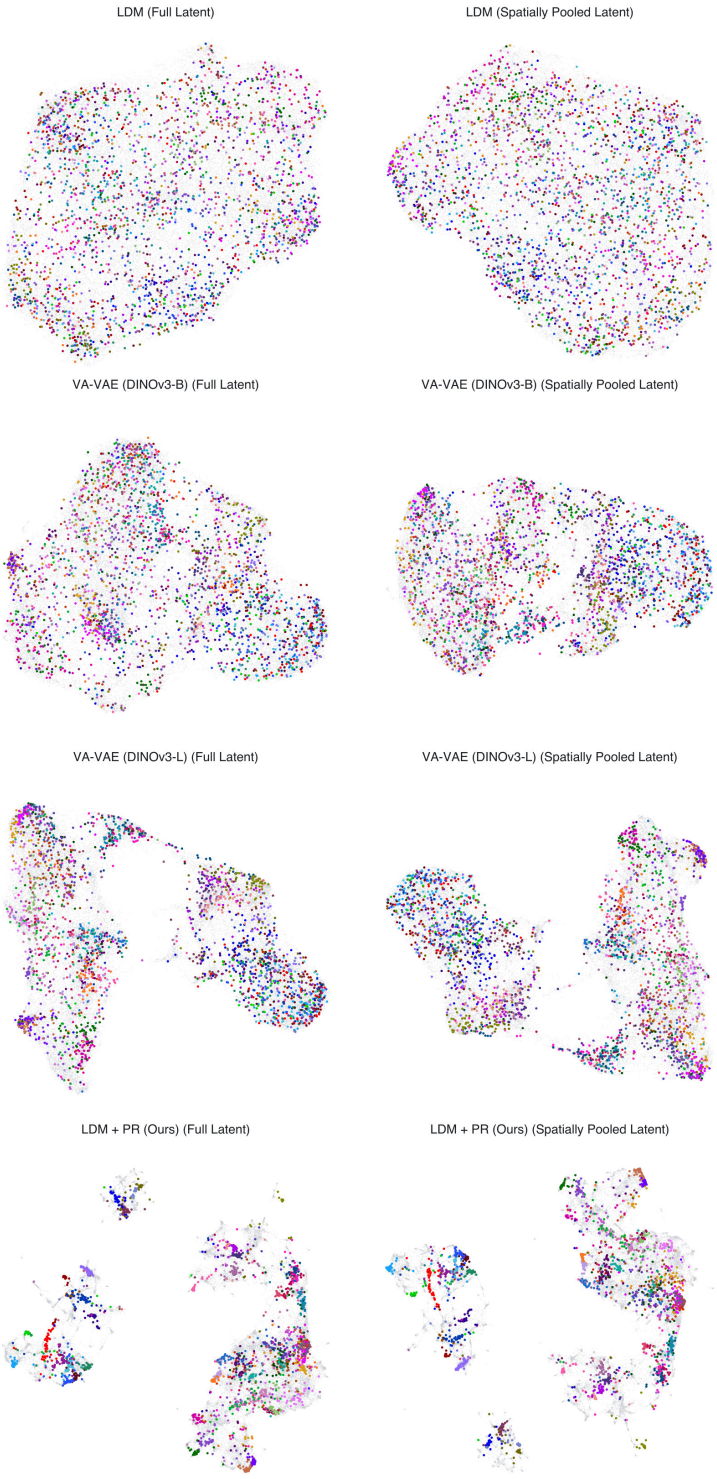


Figure 3: UMAP visualization of full and spatially pooled latents from LDM, VA-VAE, and PR (ours). We encode 50K validation images and plot latents from 50 selected classes, colored by class label. Latents are normalized per dimension before UMAP. PR produces noticeably tighter and better separated class clusters. Zoom in for the best view. Better view with colors.

B CLASSIFIER-FREE GUIDANCE VISUALIZATIONS

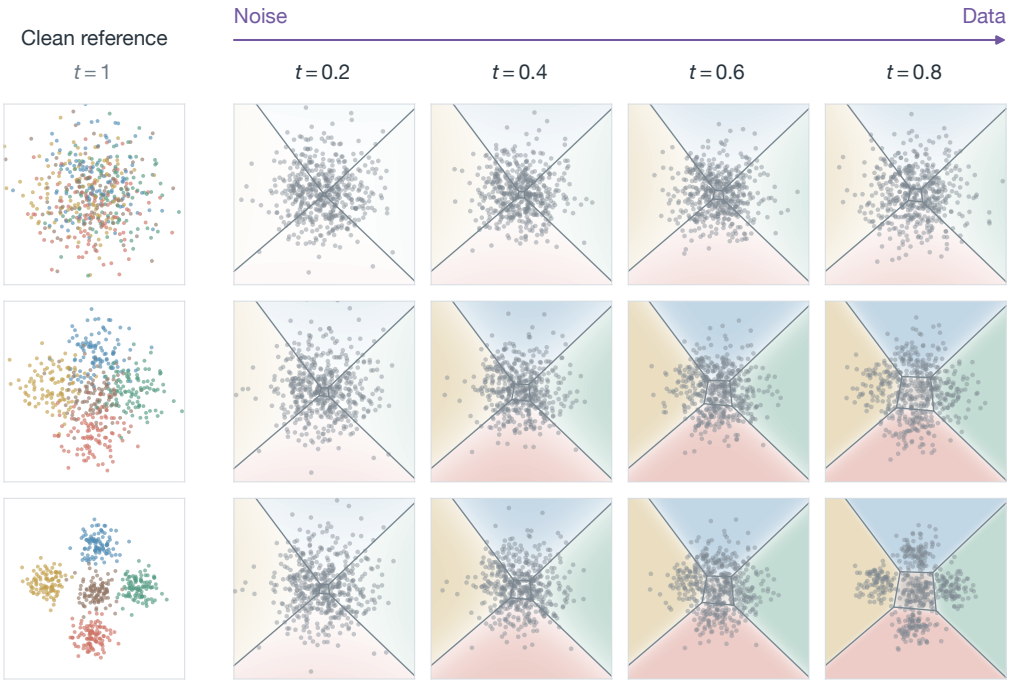


Figure 4: Noisy marginals of z_t at several $t \in (0, 1)$ for toy Gaussian mixtures with fixed total variance but different allocations between inter-class and intra-class variance. Background colors mark the Maximum A Posteriori (MAP) decision regions $\arg\max_c q_c(z_t)$, with opacity proportional to the confidence $\max_c q_c(z_t)$. With larger inter-class variance, classes become separable at smaller t, i.e., at noisier levels, and the confident MAP regions align more closely with the true class supports.

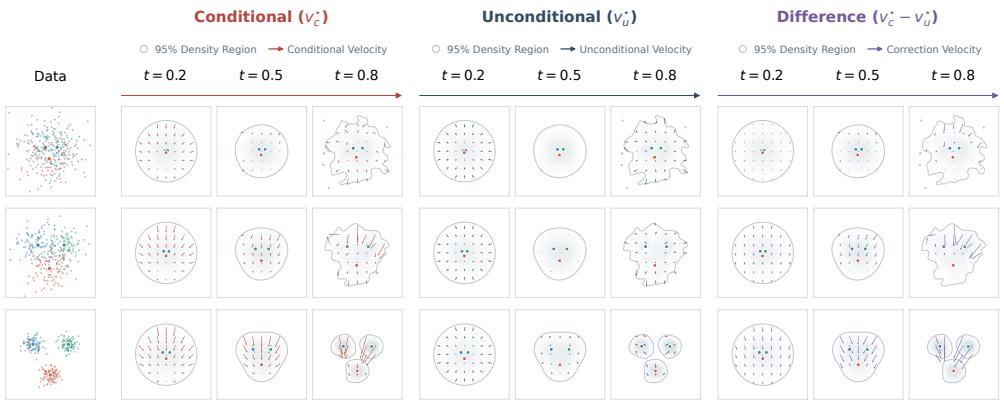


Figure 5: Oracle velocity fields on toy Gaussian mixtures with fixed total variance. From top to bottom, the fraction of variance allocated to class means increases. For each space, we show the conditional velocity $v_c^\star$ (conditioned on the red class), the unconditional velocity $v_u^\star$, and the correction $g_c^\star = v_c^\star - v_u^\star$ at $t \in \{0.2, 0.5, 0.8\}$. Gray contours mark the 95% density region of z_t, and dots mark the class means of z_t. With clearer class geometry, the conditional and correction velocities point more consistently toward the target class mean, and the correction vanishes where the posterior concentrates on the target class. Zoom in for the best view. Better view with colors.

C EXPERIMENT CONFIGURATIONS

Table 6: Training and evaluation configurations for ImageNet experiments. Full-budget experiments use 8 GPUs; reduced-budget ablations use 2 GPUs. Both budgets use the same architectures and evaluation protocol, with experiment-specific CFG search ranges listed below.

Configuration	Full Budget	Reduced Budget
Data and Architecture		
Dataset	ImageNet, 256×256	
GPUs per Run	8	2
Stage 1 Architecture	KL-VAE, downsampling factor 16	
VAE Base Channels / Multipliers	128 / $(1, 1, 2, 2, 4)$	
VAE Residual Blocks / Attention Resolution	2 per level / 16×16	
Latent Shape ($C \times H \times W$)	$32 \times 16 \times 16$	
Stage 2 Architecture	LightningDiT-XL/1	
DiT Depth / Hidden Dimension / Heads	28 / 1152 / 16	
DiT Patch Size	1×1	
DiT Components	QK-Norm, SwiGLU, RoPE, RMSNorm	
Stage 1 Training		
Image Preprocessing	Shortest-side resize to 256, random crop, no flip	
Base Objective	L1 + LPIPS + KL + adaptive adversarial loss	
KL Weight / Adversarial Weight Factor	10^{-6} / 0.5	
Discriminator Start	2,500 steps	
Optimizer (Generator and Discriminator)	Adam, $(\beta_1, \beta_2) = (0.5, 0.9)$	
Learning Rate / Schedule	10^{-4} / constant, no warmup	
Batch Size per GPU	32	32
Global Batch Size	256	64
Training Epochs	20	10
Recorded Training Steps	100,080	200,180
Stage 2 Training		
Latent Extraction	Frozen EMA encoder, posterior mode, center crop	
Latent Normalization	Per-channel mean and standard deviation	
Objective / Prediction	Linear-path flow matching / velocity	
Training Time Sampler	Logit-normal	
Optimizer	AdamW	
Optimizer (β_1, β_2)	$(0.9, 0.95)$	$(0.9, 0.995)$
Learning Rate / Schedule	10^{-4} / constant, no warmup	
Weight Decay / Max Gradient Norm	0 / 1.0	
Batch Size per GPU	128	128
Global Batch Size	1,024	256
Training Duration	80 epochs	$\sim$30 epochs
Recorded Training Steps	100,080	150,000
Condition Dropout	0.1	
Shared Training Settings		
Precision	BF16 mixed precision	
EMA Decay	0.9999	
Training Seed	42	
Sampling and Evaluation		
Evaluation Weights	EMA	
Stage 1 Evaluation	50K validation images; rFID, PSNR, LPIPS, L1	
ODE Solver / Steps	Euler / 50	
Sampling Time Grid	Uniform, no timestep shift	
CFG Interval	Enabled for $t \geq 0.11$ ($t = 0$ denotes noise)	
CFG Selection	Minimum torch-FID over 5K generated samples	
Default CFG Grid	$\{1.25, 1.50, \ldots, 3.25\}$	
Spatial-Prototype Ablation CFG Grid	–	$\{1.5, 2.0, \ldots, 7.0\}$
Final Generation Evaluation	50K samples with selected CFG; 50K without CFG	
Generation Metrics / Evaluator	FID, sFID, IS, Precision, Recall / OpenAI ADM	
Default FID Reference	ImageNet training reference batch	
Additional T2I Evaluation	Validation-reference FID and image–prompt CLIP cosine	

D Additional Experiment Results

D.1 Ablation on Classifier Type in PR

Table 7: **Ablation on the classifier type in PR (reduced compute budget).** We vary the classifier head (cosine vs. linear) and the latent on which PR is applied (clean latents only vs. noisy latents at uniformly sampled timesteps). The cosine head on clean-only latents achieves the best rFID and gFID, both with and without CFG. Applying PR on uniform-time noisy latents over-regularizes the latent space: the top PCA component degenerates into a near-constant channel that fails to capture spatial variations (i.e., over-smoothed latents). The linear head, optimized with additional weight decay, collapses to a trivial solution that arbitrarily inflates the latent norm to enlarge the logits.

Method	Stage 1			Stage 2	
	rFID ↓	PSNR ↑	LPIPS ↓	gFID ↓ w/ CFG	gFID ↓ w/o CFG
LDM	0.285	**27.13**	**0.100**	4.56	18.79
LDM + PR (Cosine Head, Clean-only Latent)	**0.279**	26.75	0.103	**3.24**	**17.11**
LDM + PR (Cosine Head, Uniform-time Noisy Latent)	0.457	26.83	0.110	14.02	41.63
LDM + PR (Linear Head, Clean-only Latent)	0.319	26.64	0.108	21.55	64.18

D.2 Ablation on Prototype Resolution

Table 8: **Ablation on prototype resolution (reduced compute budget).** All variants use the default PR setting (cosine prototype classifier on clean-only latents) and differ only in the output resolution of spatial average pooling applied before the classifier, where 16×16 corresponds to no pooling. Low-resolution prototypes perform best: 1×1 prototypes (i.e., global average pooling) achieve the best rFID and gFID without CFG, and 2×2 prototypes achieve the best gFID with CFG, while further increasing prototype resolution consistently degrades generation quality. This suggests that reducing the spatial resolution of prototypes benefits generation, likely because it avoids imposing class-level layout constraints on the latent space.

Method	Stage 1			Stage 2	
	rFID ↓	PSNR ↑	LPIPS ↓	gFID ↓ w/ CFG	gFID ↓ w/o CFG
LDM	0.285	**27.13**	**0.100**	4.56	18.79
LDM + PR (1×1 Prototypes)	**0.279**	26.75	0.103	3.24	**17.11**
LDM + PR (2×2 Prototypes)	0.312	26.41	0.109	**3.19**	20.66
LDM + PR (4×4 Prototypes)	0.307	26.40	0.109	4.04	22.53
LDM + PR (16×16 Prototypes)	0.305	26.34	0.109	15.01	38.74

D.3 Ablation on Prototype Temperature

Table 9: **Ablation on the temperature of PR (reduced compute budget).** τ scales the cosine-similarity logits of the prototype classifier. The default $\tau = 0.1$ was fixed prior to this ablation and is used throughout the main text. All PR variants improve gFID over LDM both with and without CFG, indicating that PR is robust to τ over a $10\times$ range. A smaller τ sharpens the softmax and concentrates gradients on the hardest negative prototypes, whereas a larger τ keeps the softmax from saturating and maintains a persistent pull of each latent toward its class prototype. The best gFID at $\tau = 0.3$ thus suggests that CFG benefits more from intra-class compactness than from hard-negative separation, and that the main results, obtained with an untuned τ, are conservative.

Method	**Stage 1**			**Stage 2**	
	rFID ↓	PSNR ↑	LPIPS ↓	gFID ↓ w/ CFG	gFID ↓ w/o CFG
LDM	0.285	**27.13**	**0.100**	4.56	18.79
LDM + PR ($\tau = 0.03$)	0.299	26.76	0.104	3.27	17.61
LDM + PR ($\tau = 0.1$)	**0.279**	26.75	0.103	3.24	17.11
LDM + PR ($\tau = 0.3$)	0.291	26.76	0.103	**3.14**	**16.24**

D.4 Ablation on Pseudo-label Granularity

Table 10: **Ablation on DINO pseudo-label granularity (reduced compute budget).** This group of experiments uses a separate hyper-parameter setup, so the results are comparable within this table but not with other tables. Reconstruction is largely insensitive to the choice of targets. Unguided gFID degrades as K increases, while guided gFID remains comparable across K except for overly fine clusters. The nested variant (i.e., partitioning each of the 64 clusters into 16 sub-clusters) achieves the best generation among unsupervised targets (best unguided gFID and guided gFID close to GT labels), showing the feasibility of hierarchical PR.

Labels / Clusters for PR	**Stage 1**			**Stage 2**	
	rFID ↓	PSNR ↑	LPIPS ↓	gFID ↓ w/ CFG	gFID ↓ w/o CFG
DINOv3-B $K = 64$	0.264	**28.09**	0.096	3.01	14.07
DINOv3-B $K = 256$	**0.254**	28.00	**0.095**	3.04	14.99
DINOv3-B $K = 1000$	0.294	27.59	0.101	2.99	17.60
DINOv3-B $K = 4096$	0.263	27.99	0.095	3.27	17.93
DINOv3-B $K = 64 \times 16$ (Nested)	0.264	27.99	0.096	2.94	**13.63**
GT Labels ($K = 1000$)	0.263	27.82	0.098	**2.87**	14.60

D.5 Composability of PR and VA-VAE

Table 11: **Composability of PR and VA-VAE.** This group of experiments uses the same separate hyper-parameter setup as Tab. 10, so the results are comparable with that table but not with the others. We apply PR on top of VA-VAE, which aligns to DINOv3-L features. Combining the two yields better generation both with and without CFG, at the cost of mildly degraded reconstruction, indicating that PR is complementary to external feature alignment for generation.

Method	Stage 1			Stage 2	
	rFID ↓	PSNR ↑	LPIPS ↓	gFID ↓ w/ CFG	gFID ↓ w/o CFG
LDM	0.248	28.21	0.095	4.25	18.85
LDM + PR	0.264	27.82	0.098	2.87	14.60
VAVAE (DINOv3-L)	**0.229**	**28.27**	**0.092**	3.04	12.71
VAVAE (DINOv3-L) + PR	0.282	27.55	0.101	**2.57**	**10.89**

E Derivation of Equations and Claims in Main Text

E.1 Notation and setup

Setup. Let $C \in \{1, \dots, K\}$ denote the class label and $Z \in \mathbb{R}^d$ the clean latent, with $\mathbb{E}\|Z\|^2 < \infty$. Let $\epsilon \sim \mathcal{N}(0, I)$ be independent of (Z, C). For a fixed $t \in (0, 1)$, define $Z_t = tZ + (1-t)\epsilon$ and $v = Z - \epsilon$. We use z, c, and z_t for realizations, write $q_j(z_t, t) = \Pr(C = j \mid Z_t = z_t, t)$, and abbreviate it as q_j when the arguments are clear. The Bayes-optimal conditional and unconditional velocity predictors are denoted by $v_c^\star$ and $v_u^\star$, respectively, and $g_c^\star = v_c^\star - v_u^\star$. All expectations and entropies involving Z_t are evaluated at the stated fixed time unless time is explicitly averaged over. Under the Gaussian class model, $Z \mid C = j \sim \mathcal{N}(\mu_j, \sigma^2 I)$ with $\sigma^2 > 0$, and we write $s_t^2 = t^2\sigma^2 + (1-t)^2$. Class priors are fixed and positive, but need not be equal unless stated otherwise. For PR, the latent is viewed as a tensor $z \in \mathbb{R}^{d_c \times h \times w}$, with $d = d_c hw$, pooled feature $\bar{z} = \mathrm{AvgPool}(z)$, and normalized features $\tilde{z} = \bar{z}/\|\bar{z}\|$ and $\tilde{p}_j = p_j/\|p_j\|$. These normalized vectors are assumed to have nonzero denominators. We reserve q_j for the true posterior and π_j for the prototype classifier. All logarithms are natural.

E.2 Posterior-mean expression for CFG (Eq. 1)

Derivation. The interpolant gives $\epsilon = (z_t - tz)/(1-t)$, so the velocity target can be written as

$$v = z - \epsilon = \frac{z - z_t}{1-t}.$$

Under squared-error prediction, the Bayes-optimal predictors are conditional expectations. Since z_t and t are fixed within these expectations,

$$v_c^\star = \frac{\mathbb{E}[Z \mid Z_t = z_t, t, C = c] - z_t}{1-t}, \qquad v_u^\star = \frac{\mathbb{E}[Z \mid Z_t = z_t, t] - z_t}{1-t}.$$

Subtracting cancels the common observed latent and yields

$$g_c^\star(z_t, t) = \frac{\mathbb{E}[Z \mid Z_t = z_t, t, C = c] - \mathbb{E}[Z \mid Z_t = z_t, t]}{1-t}. \tag{10}$$

This identity does not require the Gaussian class model.

E.3 Gaussian class geometry of CFG (Eq. 2)

Derivation. Under the Gaussian class model, the independent within-class variation and corruption noise give

$$Z_t \mid C = j \sim \mathcal{N}(t\mu_j, s_t^2 I), \qquad s_t^2 = t^2\sigma^2 + (1-t)^2. \tag{11}$$

Bayes' rule therefore gives

$$q_j(z_t, t) = \frac{\Pr(C = j) \exp\left(-\frac{\|z_t - t\mu_j\|^2}{2s_t^2}\right)}{\sum_{k=1}^K \Pr(C = k) \exp\left(-\frac{\|z_t - t\mu_k\|^2}{2s_t^2}\right)}.$$

To compute the clean posterior mean within class j, combine the Gaussian prior and corruption likelihood:

$$p(z \mid z_t, t, C = j) \propto \exp\left(-\frac{\|z - \mu_j\|^2}{2\sigma^2} - \frac{\|z_t - tz\|^2}{2(1-t)^2}\right).$$

Completing the square gives

$$\begin{aligned}\mathbb{E}[Z \mid Z_t = z_t, t, C = j] &= \left(\frac{1}{\sigma^2} + \frac{t^2}{(1-t)^2}\right)^{-1} \left(\frac{\mu_j}{\sigma^2} + \frac{tz_t}{(1-t)^2}\right) \\ &= \frac{t\sigma^2}{s_t^2} z_t + \frac{(1-t)^2}{s_t^2} \mu_j.\end{aligned}$$

Averaging over the posterior class probabilities gives

$$\mathbb{E}[Z \mid Z_t = z_t, t] = \frac{t\sigma^2}{s_t^2} z_t + \frac{(1-t)^2}{s_t^2} \sum_{j=1}^K q_j \mu_j.$$

Substituting both posterior means into Eq. 10 yields

$$g_c^\star(z_t,t) = \frac{1-t}{s_t^2}\left(\mu_c - \sum_{j=1}^{K} q_j \mu_j\right) = \frac{1-t}{(1-t)^2 + t^2\sigma^2}\sum_{j\neq c} q_j(\mu_c - \mu_j),$$

where the last equality uses $\sum_j q_j = 1$.

E.4 Class separation under corruption (Eq. 3)

Derivation. By Eq. 11, the corrupted class means are $t\mu_c$ and $t\mu_j$, with shared covariance $s_t^2 I$. Their squared standardized distance is therefore

$$\delta_{cj}^2(t) = (t\mu_c - t\mu_j)^\top (s_t^2 I)^{-1}(t\mu_c - t\mu_j) = \frac{t^2\|\mu_c - \mu_j\|^2}{t^2\sigma^2 + (1-t)^2}.$$

E.5 Fixed-variance construction (Eq. 4)

Verification. Let $\mu_c = \sqrt{\alpha}\,\mu_c^{\mathrm{ref}}$ and $\sigma^2 = 1-\alpha$, where $\mathbb{E}_C[\mu_C^{\mathrm{ref}}] = 0$ and $\mathbb{E}_C\|\mu_C^{\mathrm{ref}}\|^2 = d$. Then $\mathbb{E}[Z] = 0$, and the inter-class and intra-class variance contributions are, respectively,

$$\mathbb{E}_C\|\mathbb{E}[Z \mid C] - \mathbb{E}[Z]\|^2 = \alpha\,\mathbb{E}_C\|\mu_C^{\mathrm{ref}}\|^2 = \alpha d,$$
$$\mathbb{E}_C \operatorname{tr}\operatorname{Cov}(Z \mid C) = d\sigma^2 = (1-\alpha)d.$$

The law of total variance therefore gives

$$\operatorname{tr}\operatorname{Cov}(Z) = \alpha d + (1-\alpha)d = d.$$

Thus, the construction reallocates a fixed total variance between class means and within-class variation.

E.6 Posterior cross-entropy and conditional entropy (Eq. 5)

Derivation. At fixed t, sampling (Z, C, ϵ) induces the joint distribution of (Z_t, C). Conditioning first on Z_t and using the true class posterior gives

$$\begin{aligned}\mathcal{L}_{\mathrm{post}}(t) &= \mathbb{E}_{Z,C,\epsilon}[-\log q_C(tZ + (1-t)\epsilon, t)] \\ &= \mathbb{E}_{Z_t}\left[\mathbb{E}_{C\mid Z_t,t}[-\log q_C(Z_t,t)]\right] \\ &= \mathbb{E}_{Z_t}\left[-\sum_{j=1}^{K} q_j(Z_t,t)\log q_j(Z_t,t)\right] \\ &= H(C \mid Z_t).\end{aligned}$$

The equality relies on using the true posterior, rather than an arbitrary classifier.

E.7 Posterior gradient and CFG (Eq. 6)

Derivation. The corruption likelihood is $p(z_t \mid z, t) = \mathcal{N}(z_t; tz, (1-t)^2 I)$. Differentiating its marginal with respect to z_t gives

$$\begin{aligned}\nabla_{z_t}\log p(z_t \mid C=c, t) &= \frac{\mathbb{E}_{Z\mid C=c}[\nabla_{z_t} p(z_t \mid Z, t)]}{p(z_t \mid C=c, t)} \\ &= \mathbb{E}\left[\frac{tZ - z_t}{(1-t)^2}\,\middle|\, Z_t = z_t, t, C=c\right] \\ &= \frac{t\,\mathbb{E}[Z \mid Z_t = z_t, t, C=c] - z_t}{(1-t)^2}.\end{aligned}$$

The same calculation without conditioning on the class yields

$$\nabla_{z_t} \log p(z_t \mid t) = \frac{t\,\mathbb{E}[Z \mid Z_t = z_t, t] - z_t}{(1-t)^2}.$$

By Bayes' rule, the class prior contributes no derivative with respect to z_t. Subtracting the two scores and using Eq. 10, we obtain

$$\begin{aligned}\nabla_{z_t} \log q_c(z_t, t) &= \nabla_{z_t} \log p(z_t \mid C = c, t) - \nabla_{z_t} \log p(z_t \mid t) \\ &= \frac{t}{(1-t)^2}\left(\mathbb{E}[Z \mid Z_t = z_t, t, C = c] - \mathbb{E}[Z \mid Z_t = z_t, t]\right) \\ &= \frac{t}{1-t} g_c^\star(z_t, t).\end{aligned}$$

Finally, hold q_c, t, and ϵ fixed and differentiate through $z_t = tz + (1-t)\epsilon$:

$$\nabla_z \log q_c(tz + (1-t)\epsilon, t) = t\nabla_{z_t} \log q_c(z_t, t) = \frac{t^2}{1-t} g_c^\star(z_t, t).$$

This is the per-sample negative cross-entropy gradient with a fixed posterior function; it does not include changes in that function induced by changing the underlying latent distribution. The identity does not require the Gaussian class model.

E.8 Monotonic posterior loss under variance allocation (Eqs. 4 and 5)

Derivation. Under the Gaussian family in Appendix E.5, write $Z = \sqrt{\alpha}\,\mu_C^{\mathrm{ref}} + \sqrt{1-\alpha}\,\eta$, where $\eta \sim \mathcal{N}(0, I)$ is independent of C and ϵ. The corrupted latent is then

$$Z_t = t\sqrt{\alpha}\,\mu_C^{\mathrm{ref}} + t\sqrt{1-\alpha}\,\eta + (1-t)\epsilon.$$

Combining the two Gaussian noise terms and rescaling the observation gives the equivalent channel

$$Y_\alpha := \frac{Z_t}{\sqrt{t^2(1-\alpha) + (1-t)^2}} = \sqrt{\gamma_t(\alpha)}\,\mu_C^{\mathrm{ref}} + \xi, \qquad \gamma_t(\alpha) := \frac{t^2\alpha}{t^2(1-\alpha) + (1-t)^2},$$

where $\xi \sim \mathcal{N}(0, I)$ is independent of C. The rescaling is invertible, so $H(C \mid Z_t) = H(C \mid Y_\alpha)$. For every fixed $t \in (0, 1)$,

$$\frac{d\gamma_t(\alpha)}{d\alpha} = \frac{t^2\left(t^2 + (1-t)^2\right)}{\left(t^2(1-\alpha) + (1-t)^2\right)^2} > 0.$$

To compare $0 < \alpha_1 < \alpha_2 < 1$, write $\gamma_i = \gamma_t(\alpha_i)$ and construct

$$Y_2 = \sqrt{\gamma_2}\,\mu_C^{\mathrm{ref}} + \xi_2, \qquad \widehat{Y}_1 = \sqrt{\frac{\gamma_1}{\gamma_2}}\,Y_2 + \sqrt{1 - \frac{\gamma_1}{\gamma_2}}\,\xi',$$

where ξ_2 and ξ' are independent standard Gaussian variables, both independent of C. Conditional on C, the combined noise in $\widehat{Y}_1$ has covariance I, so $(C, \widehat{Y}_1)$ has the same distribution as (C, Y_{α_1}). Moreover, $C \to Y_2 \to \widehat{Y}_1$ is a Markov chain. The data-processing inequality therefore gives

$$I(C; Y_{\alpha_1}) = I(C; \widehat{Y}_1) \le I(C; Y_2) = I(C; Y_{\alpha_2}).$$

Since the class priors and hence $H(C)$ are fixed,

$$H(C \mid Y_{\alpha_2}) = H(C) - I(C; Y_{\alpha_2}) \le H(C) - I(C; Y_{\alpha_1}) = H(C \mid Y_{\alpha_1}).$$

Using the entropy identity in Appendix E.6, we conclude that $\mathcal{L}_{\mathrm{post}}(t)$ is non-increasing in α at every fixed $t \in (0, 1)$.

E.9 PROTOTYPE POSTERIOR AND CLASS-CONTRAST GRADIENT (EQS. 7 AND 8)

Gaussian posterior form. For a clean latent under the Gaussian class model, Bayes' rule and expansion of the squared distance give

$$q_j(z,1) = \frac{\exp\left(\frac{\mu_j^\top z}{\sigma^2} - \frac{\|\mu_j\|^2}{2\sigma^2} + \log \Pr(C=j)\right)}{\sum_{k=1}^K \exp\left(\frac{\mu_k^\top z}{\sigma^2} - \frac{\|\mu_k\|^2}{2\sigma^2} + \log \Pr(C=k)\right)}.$$

With equal priors and equal-norm class means, the bias terms are common across classes and cancel:

$$q_j(z,1) = \frac{\exp(\mu_j^\top z/\sigma^2)}{\sum_{k=1}^K \exp(\mu_k^\top z/\sigma^2)}.$$

PR adopts this softmax inner-product form on normalized pooled features, with learnable normalized prototypes and temperature $\tau > 0$. Its posterior and cross-entropy are defined as

$$\pi_j(z) = \frac{\exp(\tilde{z}^\top \tilde{p}_j/\tau)}{\sum_{k=1}^K \exp(\tilde{z}^\top \tilde{p}_k/\tau)}, \qquad \mathcal{L}_{\mathrm{PR}} = -\log \pi_c(z).$$

This surrogate matches the Gaussian discriminant form without assuming that normalized pooled features themselves follow the original Gaussian class model.

Gradient derivation. Expand the cross-entropy as

$$\mathcal{L}_{\mathrm{PR}} = -\frac{\tilde{z}^\top \tilde{p}_c}{\tau} + \log \sum_{k=1}^K \exp(\tilde{z}^\top \tilde{p}_k/\tau).$$

Holding the prototypes fixed and differentiating with respect to the normalized feature argument gives

$$\begin{aligned}\nabla_{\tilde{z}} \mathcal{L}_{\mathrm{PR}} &= -\frac{\tilde{p}_c}{\tau} + \frac{\sum_{j=1}^K \exp(\tilde{z}^\top \tilde{p}_j/\tau)\, \tilde{p}_j/\tau}{\sum_{k=1}^K \exp(\tilde{z}^\top \tilde{p}_k/\tau)} \\ &= \frac{1}{\tau}\left(\sum_{j=1}^K \pi_j \tilde{p}_j - \tilde{p}_c\right).\end{aligned}$$

Using $\sum_j \pi_j = 1$, we obtain

$$\begin{aligned}-\nabla_{\tilde{z}} \mathcal{L}_{\mathrm{PR}} &= \frac{1}{\tau}\left(\tilde{p}_c - \sum_{j=1}^K \pi_j \tilde{p}_j\right) \\ &= \frac{1}{\tau}\sum_{j \neq c} \pi_j (\tilde{p}_c - \tilde{p}_j).\end{aligned}$$

This has the same posterior-weighted class-contrast structure as the expression in Appendix E.3, with normalized prototypes and classifier probabilities replacing class means and true posterior probabilities. The derivative is with respect to $\tilde{z}$ as the classifier input, before backpropagation through normalization.

F Complementary Perspectives on Classifier-Free Guidance

We examine CFG from three complementary perspectives: the irreducible uncertainty of velocity prediction, the error in a learned guidance correction, and the estimation accuracy attainable under a controlled model. The first isolates the role of latent geometry, the second separates conditional error contrasts from branch consistency, and the third connects these quantities through an explicit estimator. Throughout, $Z_t = tZ + (1-t)\epsilon$, $v = Z - \epsilon$, and $t \in (0,1)$ is fixed, with $\epsilon \sim \mathcal{N}(0, I)$ independent of (Z, C). Expectations are taken over the forward process generated by (Z, C, ϵ). We use $q_j = \Pr(C = j \mid Z_t, t)$, $v_j^\star = \mathbb{E}[v \mid Z_t, t, C = j]$, $v_u^\star = \mathbb{E}[v \mid Z_t, t]$, and $g_c^\star = v_c^\star - v_u^\star$.

F.1 Class Geometry and Irreducible Velocity Uncertainty

Takeaway and setup. Under the Gaussian class model, reallocating a fixed total variance toward class means strictly reduces the conditional velocity Bayes risk and does not increase the unconditional risk. Assume $Z \mid C = c \sim \mathcal{N}(\mu_c, \sigma^2 I)$ and write $s_t^2 = (1-t)^2 + t^2\sigma^2$. For the geometric comparison, fix the class priors and reference means satisfying $\mathbb{E}_C[\mu_C^{\mathrm{ref}}] = 0$ and $\mathbb{E}_C\|\mu_C^{\mathrm{ref}}\|^2 = d$, and set $\mu_c = \sqrt{\alpha}\,\mu_c^{\mathrm{ref}}$ and $\sigma^2 = 1 - \alpha$, with $\alpha \in (0,1)$.

Velocity Bayes risks. The conditional and unconditional irreducible errors satisfy

$$\begin{aligned} \mathcal{R}_{\mathrm{cond}}(t) &:= \mathbb{E}\|v - v_C^\star\|^2 = \frac{d\sigma^2}{s_t^2}, \\ \mathcal{R}_{\mathrm{uncond}}(t) &:= \mathbb{E}\|v - v_u^\star\|^2 = \mathcal{R}_{\mathrm{cond}}(t) + \mathbb{E}\|g_C^\star\|^2. \end{aligned} \tag{12}$$

To obtain the first identity, Gaussian conditioning gives

$$\mathrm{Cov}(Z \mid Z_t, t, C) = \left(\frac{1}{\sigma^2} + \frac{t^2}{(1-t)^2}\right)^{-1} I = \frac{\sigma^2(1-t)^2}{s_t^2} I.$$

Since $v = (Z - Z_t)/(1-t)$, its conditional covariance is $\sigma^2 I / s_t^2$, whose trace gives $\mathcal{R}_{\mathrm{cond}}$. For the second identity, write $v - v_u^\star = (v - v_C^\star) + g_C^\star$. The cross term vanishes because $\mathbb{E}[v - v_C^\star \mid Z_t, t, C] = 0$.

Effect of variance allocation. For any $0 < \alpha_1 < \alpha_2 < 1$, we have

$$\begin{aligned} \mathcal{R}_{\mathrm{cond}}(t; \alpha_2) &< \mathcal{R}_{\mathrm{cond}}(t; \alpha_1), \\ \mathcal{R}_{\mathrm{uncond}}(t; \alpha_2) &\le \mathcal{R}_{\mathrm{uncond}}(t; \alpha_1). \end{aligned} \tag{13}$$

The conditional statement follows directly from

$$\frac{\partial \mathcal{R}_{\mathrm{cond}}(t;\alpha)}{\partial \alpha} = -\frac{d(1-t)^2}{\big((1-t)^2 + t^2(1-\alpha)\big)^2} < 0.$$

For the unconditional statement, couple the two models as follows. Let Z_2 follow the family at α_2, let $\epsilon_2 \sim \mathcal{N}(0, I)$ be independent, and take independent standard Gaussians η_0, ϵ_0 independent of (Z_2, C, ϵ_2). Set

$$\rho = \sqrt{\frac{\alpha_1}{\alpha_2}}, \qquad Z_1 = \rho Z_2 + \sqrt{1-\rho^2}\,\eta_0, \qquad \epsilon_1 = \rho\epsilon_2 + \sqrt{1-\rho^2}\,\epsilon_0.$$

Then $Z_1 \mid C = c \sim \mathcal{N}(\sqrt{\alpha_1}\mu_c^{\mathrm{ref}}, (1-\alpha_1)I)$, and ϵ_1 is independent of (Z_1, C). Write $Y_i = tZ_i + (1-t)\epsilon_i$ and $V_i = Z_i - \epsilon_i$ for $i = 1, 2$. Remove a common linear predictor by defining

$$A_t = t^2 + (1-t)^2, \qquad \kappa_t = \frac{2t-1}{A_t}, \qquad r_i = V_i - \kappa_t Y_i.$$

The fixed total variance implies

$$\mathbb{E}\|Y_i\|^2 = dA_t, \qquad \mathbb{E}[V_i^\top Y_i] = d(2t-1), \qquad \mathbb{E}\|V_i\|^2 = 2d,$$

and hence $\mathbb{E}\|r_i\|^2 = d/A_t$ for both models. Moreover,

$$Y_1 = \rho Y_2 + \sqrt{1-\rho^2}\,Y_0, \qquad r_1 = \rho r_2 + \sqrt{1-\rho^2}\,r_0,$$

where $Y_0 = t\eta_0 + (1-t)\epsilon_0$ and $r_0 = \eta_0 - \epsilon_0 - \kappa_t Y_0$. These two auxiliary variables are jointly Gaussian with zero cross-covariance, so r_0 is independent of Y_0, and both are independent of (Y_2, r_2). Therefore,

$$\mathbb{E}[r_1 \mid Y_1] = \rho\, \mathbb{E}[\mathbb{E}[r_2 \mid Y_2] \mid Y_1], \qquad \mathbb{E}\|\mathbb{E}[r_1 \mid Y_1]\|^2 \le \rho^2 \mathbb{E}\|\mathbb{E}[r_2 \mid Y_2]\|^2.$$

Finally, subtracting a known function of the observation does not change the Bayes risk. Orthogonality of conditional expectation gives

$$\begin{aligned}
\mathcal{R}_{\mathrm{uncond}}(t;\alpha_1) &= \frac{d}{A_t} - \mathbb{E}\|\mathbb{E}[r_1 \mid Y_1]\|^2 \\
&\ge \frac{d}{A_t} - \rho^2 \mathbb{E}\|\mathbb{E}[r_2 \mid Y_2]\|^2 \\
&\ge \frac{d}{A_t} - \mathbb{E}\|\mathbb{E}[r_2 \mid Y_2]\|^2 = \mathcal{R}_{\mathrm{uncond}}(t;\alpha_2).
\end{aligned}$$

Connection to diffusability. Velocity irreducible variance has been studied as a statistical indicator of latent diffusability (Zhong et al., 2026). The conditional risk in Eq. 12 is the isotropic Gaussian instance of that quantity. Equation 13 connects this perspective to class geometry: at a fixed latent scale, increasing the inter-class variance fraction reduces velocity-target uncertainty in the conditional branch and preserves or reduces it in the unconditional branch.

F.2 Conditional Error Contrasts and Branch Consistency

Takeaway and setup. The error in a learned CFG correction separates exactly into differences between conditional prediction errors and a consistency term relating the two branches. This decomposition holds without Gaussian assumptions. Let $e_j = \widehat{v}_j - v_j^\star$ and $e_u = \widehat{v}_u - v_u^\star$, and define $\widehat{g}_c = \widehat{v}_c - \widehat{v}_u$. Treat the learned predictors as fixed functions when averaging over the forward distribution.

Exact correction-error decomposition. The expected correction error is

$$\begin{aligned} \mathcal{E}_g(t) &:= \mathbb{E}\|\widehat{g}_C - g_C^\star\|^2 = \mathcal{E}_g^{(1)}(t) + \mathcal{E}_g^{(2)}(t), \\ \mathcal{E}_g^{(1)}(t) &:= \frac{1}{2}\mathbb{E}\sum_{i,j} q_i q_j \|e_i - e_j\|^2, \\ \mathcal{E}_g^{(2)}(t) &:= \mathbb{E}\left\|\widehat{v}_u - \sum_j q_j \widehat{v}_j\right\|^2. \end{aligned} \tag{14}$$

To see this, the tower property gives $v_u^\star = \sum_j q_j v_j^\star$. Define $\bar{e} = \sum_j q_j e_j$ and $\Delta_u = \widehat{v}_u - \sum_j q_j \widehat{v}_j = e_u - \bar{e}$. Then

$$\widehat{g}_C - g_C^\star = e_C - e_u = (e_C - \bar{e}) - \Delta_u.$$

Conditional on Z_t, we have $\mathbb{E}[e_C - \bar{e} \mid Z_t, t] = 0$, so the cross term vanishes. The remaining conditional variance satisfies

$$\sum_j q_j \|e_j - \bar{e}\|^2 = \sum_j q_j \|e_j\|^2 - \|\bar{e}\|^2 = \frac{1}{2}\sum_{i,j} q_i q_j \|e_i - e_j\|^2,$$

which proves Eq. 14. The first term measures error differences across jointly plausible classes; any component shared by all e_j cancels. The second measures agreement between the learned unconditional prediction and the posterior-weighted conditional predictions.

Posterior weighting and prediction uncertainty. Define the expected posterior ambiguity as

$$U_t := \mathbb{E}\left[1 - \sum_j q_j^2\right] = \mathbb{E}\sum_{i \neq j} q_i q_j.$$

If $\|e_i - e_j\| \leq B$ uniformly, then $\mathcal{E}_g^{(1)}(t) \leq B^2 U_t/2$. Thus, for a common error-difference bound, lower posterior ambiguity tightens the bound on the conditional contrast term. These learned errors are distinct from the irreducible uncertainty in Appendix F.1; for example, conditional prediction obeys

$$\mathbb{E}\|v - \widehat{v}_C\|^2 = \mathcal{R}_{\mathrm{cond}}(t) + \mathbb{E}\|e_C\|^2.$$

Here the first term is determined by the data distribution, while the second measures approximation of its oracle predictor.

F.3 Geometry-Dependent Estimation Error under Consistent Branches

Takeaway and setup. Under a controlled Gaussian estimator, within-class compactness and low posterior ambiguity jointly reduce correction and guided-field estimation errors. Use the same fixed-variance family as in Appendix F.1. Assume that the common linear term of the conditional predictors is known, estimate each class offset from n independent residual samples, and construct the unconditional estimate using the true posterior weights. Estimation samples are independent across classes and independent of the test sample. Expectations below also average over these estimation samples, with n and the guidance scale $\omega \geq 1$ held fixed across the family.

Estimating the class offsets. The Gaussian conditional predictor has the form

$$v_j^\star(z_t, t) = a_t z_t + h_j, \qquad a_t = \frac{t\sigma^2 - (1-t)}{s_t^2}, \qquad h_j = \frac{1-t}{s_t^2}\mu_j.$$

For class j, a velocity residual $v - a_t Z_t$ has mean h_j and covariance $\sigma^2 I / s_t^2$: subtracting h_j gives $v - v_j^\star$, whose conditional mean is zero and whose conditional covariance was computed in Appendix F.1. Estimate

$$\widehat{h}_j = \frac{1}{n}\sum_{\ell=1}^{n}\big(v_{j\ell} - a_t Z_{t,j\ell}\big), \qquad \widehat{v}_j(z_t, t) = a_t z_t + \widehat{h}_j, \qquad \widehat{v}_u = \sum_j q_j \widehat{v}_j.$$

Consequently, $e_j = \widehat{h}_j - h_j$ is independent of the test sample and satisfies

$$\mathbb{E}[e_j] = 0, \qquad \mathbb{E}\|e_j\|^2 = \frac{d\sigma^2}{n s_t^2} = \frac{\mathcal{R}_{\mathrm{cond}}(t)}{n}, \qquad \mathbb{E}[e_i^\top e_j] = 0 \quad (i \neq j).$$

Correction and guided-field errors. The construction makes $\mathcal{E}_g^{(2)}(t) = 0$ in Eq. 14. Since $\mathbb{E}\|e_i - e_j\|^2 = 2\mathcal{R}_{\mathrm{cond}}(t)/n$ for $i \neq j$, its first term gives the correction error below. For guidance, define $v_{\mathrm{cfg}}^\star = v_u^\star + \omega(v_C^\star - v_u^\star)$ and $\widehat{v}_{\mathrm{cfg}} = \widehat{v}_u + \omega(\widehat{v}_C - \widehat{v}_u)$. Then

$$\widehat{v}_{\mathrm{cfg}} - v_{\mathrm{cfg}}^\star = \omega e_C + (1-\omega)\bar{e} = \sum_j \big(\omega \mathbf{1}_{\{j=C\}} + (1-\omega)q_j\big)e_j.$$

Independence and zero-mean estimation errors remove cross terms. Averaging the squared coefficients over $C \mid Z_t$ gives

$$\sum_c q_c \sum_j \big(\omega \mathbf{1}_{\{j=c\}} + (1-\omega)q_j\big)^2 = \omega^2 + (1-\omega^2)\sum_j q_j^2.$$

Thus, the two estimation errors are

$$\begin{aligned} \mathcal{E}_g(t) &= \frac{\mathcal{R}_{\mathrm{cond}}(t)}{n}\, U_t, \\ \mathcal{E}_\omega(t) &:= \mathbb{E}\|\widehat{v}_{\mathrm{cfg}} - v_{\mathrm{cfg}}^\star\|^2 = \frac{\mathcal{R}_{\mathrm{cond}}(t)}{n}\big[1 + (\omega^2 - 1)U_t\big]. \end{aligned} \tag{15}$$

Monotonicity under variance allocation. It remains to show that U_t is non-increasing in α. Rescaling the noisy latent gives an equivalent observation $Y_\alpha = \sqrt{\gamma_t(\alpha)}\mu_C^{\mathrm{ref}} + \xi$, where $\xi \sim \mathcal{N}(0, I)$ is independent of C and

$$\gamma_t(\alpha) = \frac{t^2\alpha}{t^2(1-\alpha) + (1-t)^2}.$$

This signal-to-noise ratio increases with α. For $\alpha_1 < \alpha_2$, the Gaussian channel coupling in Appendix E.8 constructs observations with the corresponding marginals and $C \to Y_{\alpha_2} \to Y_{\alpha_1}$. Writing $\mathbf{q}_i = \Pr(C = \cdot \mid Y_{\alpha_i})$ for their posterior probability vectors, the tower property and Jensen's inequality give

$$\mathbf{q}_1 = \mathbb{E}[\mathbf{q}_2 \mid Y_{\alpha_1}], \qquad \mathbb{E}\|\mathbf{q}_1\|^2 \leq \mathbb{E}\|\mathbf{q}_2\|^2,$$

and hence $U_t(\alpha_2) \leq U_t(\alpha_1)$. Together with the decrease of $\mathcal{R}_{\mathrm{cond}}$ in Eq. 13, Eq. 15 shows that both $\mathcal{E}_g(t)$ and $\mathcal{E}_\omega(t)$ are non-increasing in α. Within this estimator, allocating more variance to class means lowers the cost of estimating each conditional predictor and reduces the posterior ambiguity that weights class contrasts.