A Hidden Semantic Bottleneck in Conditional Embeddings of Diffusion Transformers

Anonymous authors
Paper under double-blind review

Abstract

Diffusion Transformers have achieved state-of-the-art performance in class-conditional and multimodal generation, yet the structure of their learned conditional embeddings remains poorly understood. In this work, we present the first systematic study of these embeddings and uncover a notable redundancy: class-conditioned embeddings exhibit extreme angular similarity, exceeding 99% on ImageNet-1K, while continuous-condition tasks such as pose-guided image generation and video-to-audio generation reach over 99.9%. We further find that semantic information is concentrated in a small subset of dimensions, with head dimensions carrying the dominant signal and tail dimensions contributing minimally. By pruning low-magnitude dimensions–removing up to two-thirds of the embedding space–we show that generation quality and fidelity remain largely unaffected, and in some cases improve. These results reveal a semantic bottleneck in Transformer-based diffusion models, providing new insights into how semantics are encoded and suggesting opportunities for more efficient conditioning mechanisms.

1 Introduction

Transformer-based diffusion models have recently emerged as state-of-the-art architectures for generative modeling tasks across diverse domains, including class-conditional image synthesis DiT (Peebles & Xie, 2023), MDT (Gao et al., 2023), SiT (Ma et al., 2024), LightningDiT (Yao et al., 2025), Model-Guidance (MG) (Tang et al., 2025), REPA (Yu et al., 2025), pose-guided person image generation (Pham et al., 2024), and video-to-audio generation (Pham et al., 2025). These models combine the expressive capacity of Transformer backbones with diffusion processes to generate high-fidelity, semantically consistent outputs. A key component of such models is the conditional embedding vector, often formed by summing class label and timestep embeddings and injected via adaptive layer normalization (AdaLN). *Yet despite their state-of-the-art performance and broad adoption, the role and internal structure of these learned conditional embeddings remain poorly understood.*

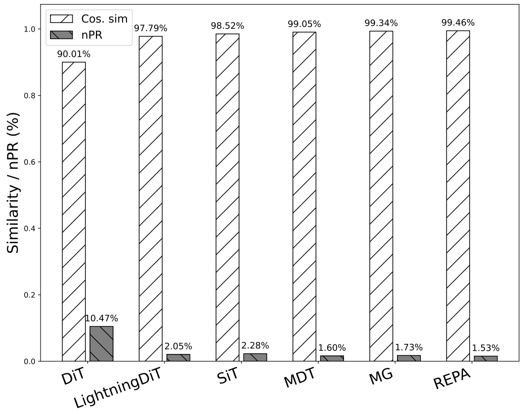


Figure 1: **Hidden Semantic Bottleneck: Extreme Alignment and Dimensional Sparsity.** Conditional vectors $\vec{c}$ in state-of-the-art Transformer diffusion models on ImageNet-1K exhibit very high pairwise cosine similarity (mostly 90–99%) while concentrating semantic information in only a few of 1,152 dimensions.

In this work, we present a systematic analysis of conditional embeddings in diffusion transformers and uncover two key findings. **(1)** Class-condition vectors exhibit extreme alignment, with cosine similarity exceeding 99% on ImageNet-1K across multiple state-of-the-art methods (Fig. 1, white bar). **(2)** The learned conditional vector $\vec{c}$ is markedly sparse: only about 10–20 of its 1,152 dimensions carry substantial magnitude, yielding a normalized participation rate (nPR) of just 1–2% (Fig. 1, gray bar). When we prune up to 66% of

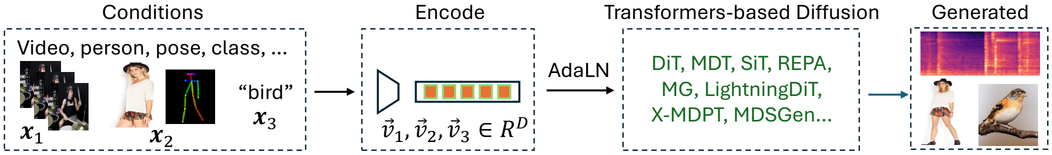


Figure 2: Transformer-based diffusion models inject conditions as a globally compact vectors $\vec{v}_i$ via AdaLN for outputs such as images or mel-spectrograms.

dimensions and perform inference with the resulting sparsified $\vec{c}$, generation quality remains essentially unchanged, exposing significant over-parameterization. These findings challenge common assumptions about semantic conditioning and indicate that diffusion transformers encode conditioning signals far more compactly than previously believed, offering a new design perspective for generative models. *Our contributions are as follows:*

- **Extreme similarity.** We present the first systematic analysis showing that, in discrete class-conditional tasks (e.g., ImageNet), transformer-based diffusion models learn class-conditioned embeddings with up to 99% pairwise cosine similarity, and in continuous-condition tasks (e.g., pose-guided image or video-conditioned audio generation), the similarity exceeds 99.9%.
- **Sparse representations.** We find that semantic information is concentrated in a small set of embedding dimensions, while most remain near zero, revealing highly sparse conditional representations.
- **Redundancy and pruning.** We demonstrate that aggressively pruning low-magnitude dimensions preserves or even improves generation quality, highlighting substantial redundancy and enabling more efficient conditioning.
- **Mechanistic insight.** We provide hypotheses, supported by analyses and theoretical reasoning, to explain the emergence of high similarity, sparsity, and pruning effectiveness.

2 Related Work

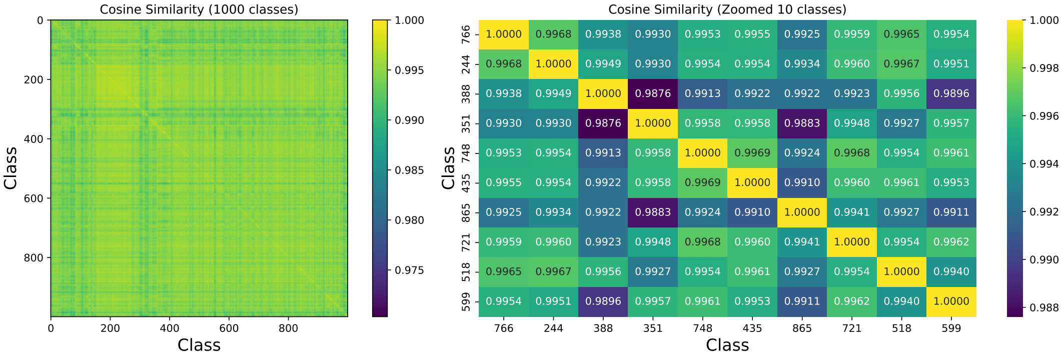
Diffusion transformers and conditioning via AdaLN. Diffusion models have progressed from U-Net backbones (Rombach et al., 2022) to transformer-based designs such as DiT (Peebles & Xie, 2023), SiT (Ma et al., 2024), LightningDiT (Yao et al., 2025), MG (Tang et al., 2025), X-MDPT (Pham et al., 2024), MDSGen (Pham et al., 2025), and UCGM (Sun et al., 2025), achieving strong results across image, audio, and multimodal generation.

These models embed conditional signals–class labels, poses, or video features–into timestep embeddings and inject them via adaptive layer normalization (AdaLN) (Fig. 2), where condition vectors modulate all layers through learned scale–shift parameters. Unlike the distributed conditioning of U-Nets, this global AdaLN mechanism motivates our study of how semantic information is encoded in transformer conditional vectors.

Prior work on conditional embedding analysis. Li et al. (2023) examined activation sparsity in Transformers for NLP and ImageNet with classification, but systematic studies of conditional embeddings in generative diffusion remain scarce. Early efforts targeted U-Net conditioning (Rombach et al., 2022; Saharia et al., 2022), while transformer-based models focused on architectural or training advances (Peebles & Xie, 2023; Ma et al., 2024; Yu et al., 2025; Tang et al., 2025; Gao et al., 2023; Pham et al., 2024; 2025). We fill this gap with an analysis of transformer conditional embeddings and their link to representation collapse.

Collapse in contrastive learning. Representation collapse–mapping diverse inputs to nearly identical embeddings–is well known in contrastive learning (Grill et al., 2020; Zbontar et al., 2021). We observe a related effect in diffusion transformers: conditional embeddings across classes reach extreme angular similarity (>99% cosine) without harming generation quality, indicating a distinct embedding usage compared to contrastive methods.

Hyperspherical embeddings and compressed codes. Our findings align with hyperspherical embedding (Liu et al., 2017) and information bottleneck theory (Tishby et al., 2000), which describe semantic compression into low-dimensional subspaces. Similar trade-offs appear in VAEs and multimodal systems (Kingma & Welling, 2013; Tsai et al., 2019). Diffu-



Class	766	244	388	351	748	435	865	721	518	599
766	1.0000	0.9968	0.9938	0.9930	0.9953	0.9955	0.9925	0.9959	0.9965	0.9954
244	0.9968	1.0000	0.9949	0.9930	0.9954	0.9954	0.9934	0.9960	0.9967	0.9951
388	0.9938	0.9949	1.0000	0.9876	0.9913	0.9922	0.9922	0.9923	0.9956	0.9896
351	0.9930	0.9930	0.9876	1.0000	0.9958	0.9958	0.9883	0.9948	0.9927	0.9957
748	0.9953	0.9954	0.9913	0.9958	1.0000	0.9969	0.9924	0.9968	0.9954	0.9961
435	0.9955	0.9954	0.9922	0.9958	0.9969	1.0000	0.9910	0.9960	0.9961	0.9953
865	0.9925	0.9934	0.9922	0.9883	0.9924	0.9910	1.0000	0.9941	0.9927	0.9911
721	0.9959	0.9960	0.9923	0.9948	0.9968	0.9960	0.9941	1.0000	0.9954	0.9962
518	0.9965	0.9967	0.9956	0.9927	0.9954	0.9961	0.9927	0.9954	1.0000	0.9940
599	0.9954	0.9951	0.9896	0.9957	0.9961	0.9953	0.9911	0.9962	0.9940	1.0000

Figure 3: Cosine similarity of conditional vectors $\vec{c} = y + t$ across 1000 ImageNet classes using REPA-XL (Yu et al., 2025). Despite distinct semantics, embeddings show over 99% similarity for nearly all class pairs. **Left**: full 1000×1000 matrix showing global alignment. **Right**: zoomed 10×10 subset for randomly chosen classes. Additional results for other SOTA methods appear in the Appendix.

sion transformers further compress conditioning into a small set of active head dimensions, leaving others largely redundant.

Conditioning injection: U-Net vs. Transformers. U-Net diffusion models inject conditions at multiple spatial scales via concatenation or cross-attention (Rombach et al., 2022; Dhariwal & Nichol, 2021), allowing localized feature extraction. Transformers, in contrast, apply global AdaLN modulation, which likely drives the observed sparsity and high similarity in conditional embeddings as semantics collapse into a few dominant dimensions.

3 Emergent Property I: Near-Uniform Cosine Similarity

3.1 Setup

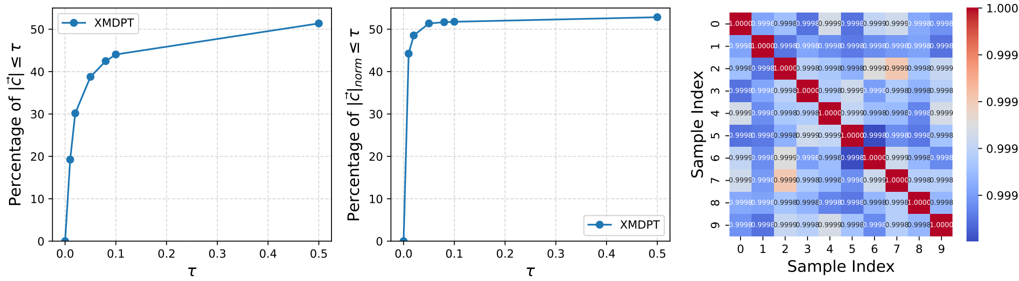
We systematically analyze six state-of-the-art diffusion transformer models–DiT (Peebles & Xie, 2023), MDT (Gao et al., 2023), SiT (Ma et al., 2024), REPA (Yu et al., 2025), LightningDiT (Yao et al., 2025), and Model-Guided (Tang et al., 2025)–using their official pretrained checkpoints released on GitHub (XL models). The primary analysis is conducted on ImageNet-1K, where we compute pairwise cosine similarity matrices across all class-conditioned vectors $\vec{c} \in \mathbb{R}^{1152}$. Each $\vec{c}$ is formed by summing the learned class embedding and timestep embedding, resulting in the final conditional vector injected into the denoising transformer backbone. To assess generality across domains, we extend the analysis to pose-guided image synthesis using X-MDPT (Pham et al., 2024) ($\vec{c} \in \mathbb{R}^{1024}$) and video-to-audio generation with MDSGen (Pham et al., 2025) ($\vec{c} \in \mathbb{R}^{768}$), again utilizing publicly available pretrained weights. This consistent evaluation setup ensures reproducibility and enables direct comparison across models and tasks.

3.2 Cosine Similarity Heatmaps

Fig. 3 (left) shows the full 1,000-class cosine-similarity matrix, where trained models reach up to $\sim$99% similarity across class pairs. For clarity, Fig. 3 (right) provides a zoomed-in view of 10 randomly sampled classes, revealing the same strong alignment. Additional results for other methods are provided in the Appendix.

3.3 Cross-Task Examination

The strong alignment of conditional embeddings extends beyond class-conditional image generation to pose-guided image synthesis and video-to-audio generation (Fig. 4). X-MDPT (Pham et al., 2024) and MDSGen (Pham et al., 2025) exhibit extreme cosine similarity—up to 99.98% on DeepFashion and 99.99% on VGGSound—even with randomly varying test samples and conditions (e.g., persons, poses, videos). Because MDSGen shows patterns nearly identical to X-MDPT, its cosine heat map is omitted. This striking consistency indicates that diverse inputs yield almost identical embeddings before denoising; we examine possible explanations in the following sections.



(a) original scale

(b) normalized scale

(c) cosine heatmap

Figure 4: **Sparsity and alignment of conditional embeddings in X-MDPT (Pham et al., 2024).** (a) and (b): With $\tau = 0.1$, over 51% of components in the conditional vectors have magnitudes below the threshold, highlighting significant sparsity. Remarkably, pruning these dimensions has minimal effect on generation quality. (c) Cosine similarity between random test samples in DeepFashion exceeds 99.9%, confirming extreme alignment across conditional embeddings.

4 Emergent Property II: Sparse Magnitude Distribution

4.1 Magnitude Histograms

Fig. 5 shows the histogram of absolute component values of $\vec{c}$: only about 10–20 of the 1,152 dimensions exceed 0.1 in magnitude, and roughly 10 exceed 1. Fig. 6 visualizes the learned conditional vector for each method, underscoring its pronounced sparsity. For completeness, we include continuous tasks such as X–MDPT and MDSGen in the Appendix; their embeddings appear less sparse, consistent with the higher participation ratio (PR) reported in Tab. 1.

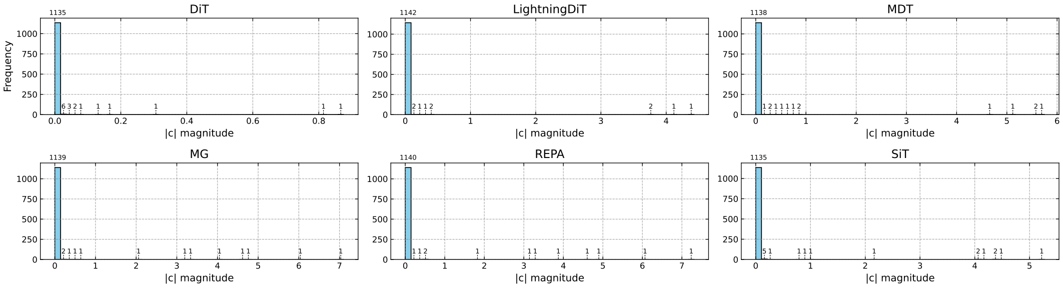


Figure 5: *Magnitude histogram distribution* of learned conditional vector embedding $\vec{c} \in \mathbb{R}^{\times 1152}$. Most dimensions have near-zero values (< 0.01), with only $\sim 5 - 20$ dimensions showing dominant magnitudes. This sparsity holds across multiple models, including DiT, MDT, LightningDiT, MG, SiT, and REPA. It is best viewed with 300% zoomed in.

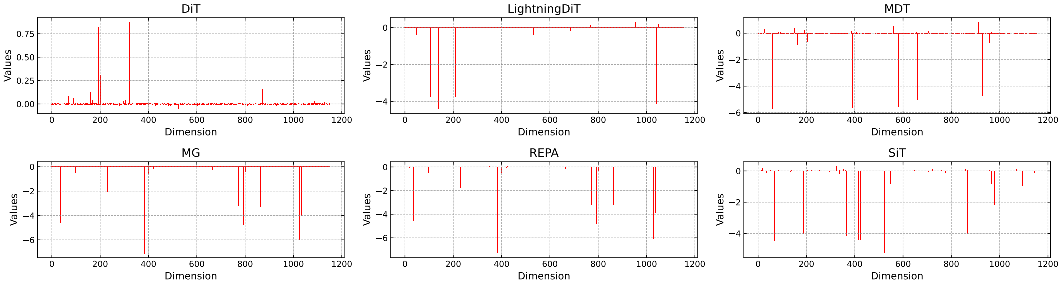


Figure 6: *Original distribution* of learned conditional vector embedding $\vec{c} \in \mathbb{R}^{1 \times 1152}$. Most dimensions have near-zero values (< 0.01), with only $\sim 5 - 20$ dimensions showing dominant magnitudes. This sparsity holds across multiple models, including DiT, MDT, LightningDiT, MG, SiT, and REPA. It is best viewed with 200% zoomed in.

Table 1: Participation Ratio (PR) in learned conditional embeddings of state-of-the-art models on Imagenet-1K generation (discrete) and DeepFashion/VGGSound (continuous).

Embedding Metrics	**DiT**	**SiT**	**MDT**	**LightningDiT**	**MG**	**REPA**	**X-MDPT**	**MDSGen**
Condition Dim (d)	1152	1152	1152	1152	1152	1152	1024	768
PR (α)	120.69	**26.25**	**18.45**	**23.70**	**19.98**	**17.67**	495.75	104.22
nPR (α_{norm})	10.47%	**2.28%**	**1.60%**	**2.05%**	**1.73%**	**1.53%**	48.42%	13.57%
Cosine Sim. (cs)	0.9001	0.9852	0.9905	0.9779	0.9934	0.9946	**0.9998**	**0.9999**

4.2 Dimension Contribution: Participation Ratio

To quantify how many dimensions effectively contribute, we compute the **participation ratio** (PR) on absolute magnitudes $v_i = |c_i|$:

$$\alpha = \mathrm{PR}(v) = \frac{\left(\sum_{i=1}^{d} v_i\right)^2}{\sum_{i=1}^{d} v_i^2}, \quad \alpha_{\text{norm}} = \frac{\alpha}{d}\ \text{, with d is dimension.} \tag{1}$$

PR estimates the number of coordinates carrying most of the total magnitude. Tab. 1 shows that state-of-the-art models (MDT, LightningDiT, MG, REPA) rely on less than 2% of dimensions, whereas continuous tasks such as X-MDPT and MDSGen use a larger fraction (13–48%) and exhibit even higher cosine similarity (up to 99.99% vs. 90–99.4%).

This suggests that continuous-condition embeddings both engage more dimensions and distribute information more uniformly, naturally leading to stronger alignment across samples compared to discrete class-conditional ImageNet generation.

5 From Observation to Action: Pruning Redundant Dimensions

Role of tail dimensions. To quantify sparsity and effective dimensionality, we define the **sparsity ratio** at threshold τ as

$$s_{\text{tail}(\tau)} = \frac{1}{d}\#\{i : |c_i| < \tau\}. \tag{2}$$

With $\tau = 0.01$, we observe a sparsity ratio of $s \approx 0.38$–0.40. We define the high-magnitude "head" as $s_{\text{head}}(\tau) = \frac{1}{d}\#\{i : |c_i| > \tau\}$ and the low-magnitude "tail" as the remaining coordinates. Using REPA as a representative model, we progressively prune $\vec{c}$ at thresholds $\tau \in \{0.01, 0.02, \dots\}$ and find that FID and CLIP scores remain stable–even after removing up to 66% of dimensions (Tab. 2), showing substantial redundancy.

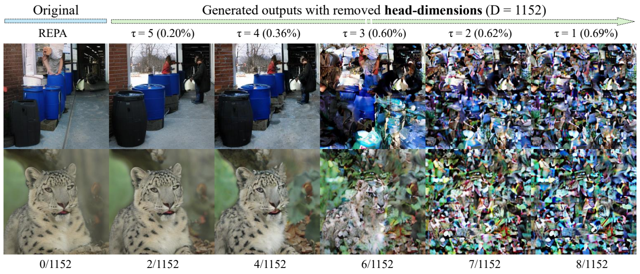


Figure 7: **Class-conditioned image generation with head removal.** ImageNet samples after pruning top-magnitude dimensions of $\vec{c}$ (threshold τ); removing only a few head dimensions markedly degrades quality.

Pruning at $\tau = 0.01$ (removing ∼38% of dimensions) preserves or improves image quality (Fig. 8, Tab. 2). Pruning during late denoising steps yields larger FID gains and modest CLIP improvements, aligning with the gradual rise in cosine similarity toward the final steps. For consistency, we report statistics using the conditional vector at the initial step t_0 and discuss possible explanations in the following sections. Next, we examine in detail how head dimensions influence generation quality and clarify how their role differs from that of the tail dimensions.

Table 2: Performance and semantic metrics under sparsification. t_i: prune every step, t_0: prune only at start, $t_{n-k,n}$: prune during last k steps.

Prune	**Threshold τ**	**# Removed Dims**	**FID ↓**	**IS ↑**	**CLIP↑**
	Baseline (REPA)	0/1152 (0%)	7.1694	**176.02**	29.746
Tail	$\tau = 0.01$ (t_i)	448/1152 (38.94%)	7.2143	171.99	29.737
	$\tau = 0.01$ (t_0)	448/1152 (38.94%)	**7.1690**	175.97	**29.807**
	$\tau = 0.01$ ($t_{n-k,n}$)	448/1152 (38.94%)	**7.1598**	175.49	**29.805**
	$\tau = 0.02$ (t_i)	762/1152 (66.21%)	9.2202	125.15	29.221
	$\tau = 0.05$ (t_i)	1110/1152 (96.41%)	56.2308	20.47	22.177
	$\tau = 5.0$ (t_i)	1149/1152 (99.80%)	356.135	1.77	21.922
Head	$\tau = 5.0$ (t_i)	2/1152 (0.20%)	7.8478	164.15	29.555
	$\tau = 1.0$ (t_i)	8/1152 (0.69%)	523.7637	1.95	22.690

Role of head dimensions. As shown in Fig. 7, removing only a few high-magnitude dimensions (e.g., 4-6/1152) dramatically degrades generation quality. In contrast, pruning up

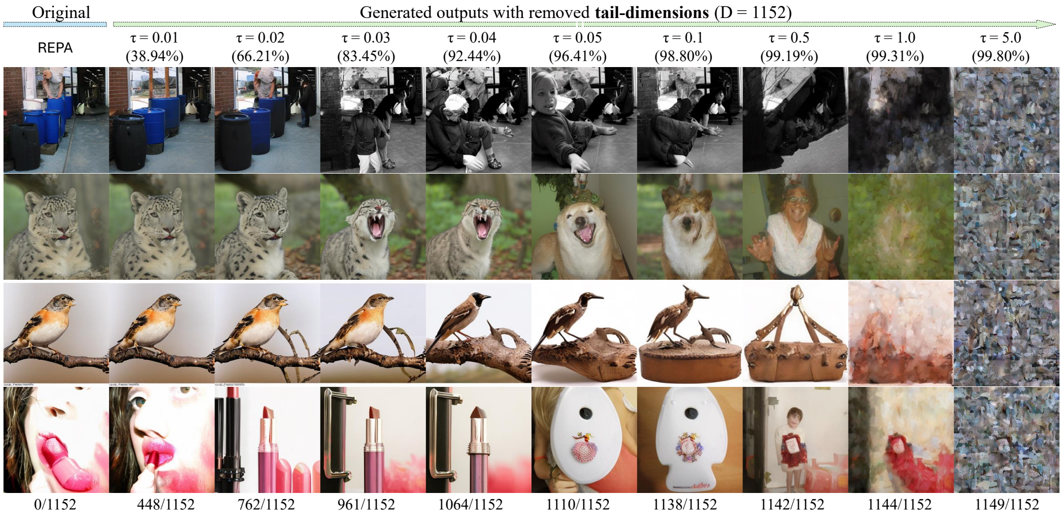
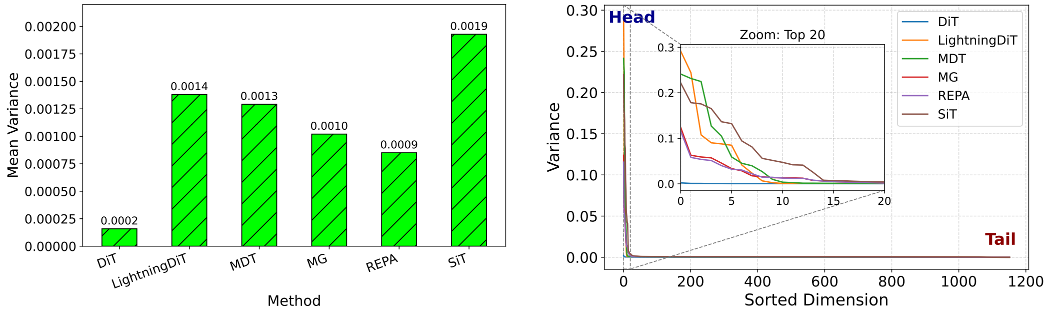


Figure 8: **Class-conditioned image generation (tail removal).** ImageNet samples with progressive removal of low-magnitude dimensions in $\vec{c}$ (threshold τ on absolute value). Image quality remains high or better baseline REPA even when 38–>80% of dimensions are pruned (generated images in the second column), as long as key head dimensions are retained.

to 66% of low-magnitude tail dimensions (762/1152) leaves quality largely intact. Variance analysis (Fig. 9) further reveals that only ∼15–20 head dimensions carry most of the variance, particularly across classes, highlighting their critical role compared to the tails.



(a) Average variance across dimensions.

(b) Variance per-dimension (sorted).

Figure 9: **Variance concentration in conditional vectors $\vec{c}$.** (a) Mean variance across models stays non-zero, showing no collapse. (b) Variance is concentrated in only 15–20 head dimensions (<2%), while the remaining 98% of tail dimensions show minimal variation, indicating that semantic information is confined to a small subspace.

Continuous task. We apply the same pruning procedure to pose-guided person image generation with X-MDPT (Pham et al., 2024) and observe consistent behavior (Fig. 10). Unlike class-conditional ImageNet, this task requires a slightly higher threshold to induce similar sparsity: $\tau = 0.1$ yields $s \approx 0.38$–0.40 (Fig. 4 a,b) while preserving generation quality (Fig. 10). More qualitative results are available in the Appendix.

6 Underlying Mechanisms Behind Similarity, Sparsity, and Pruning: Hypotheses

6.1 How Can a Model Generate Correct Outputs Despite High Similarity?

Although conditional vectors exhibit high pairwise cosine similarity, our variance analysis (Fig. 9a) shows no embedding collapse: the mean variance across models remains small but clearly non-zero (0.0002–0.0019). This contrasts with the feature collapse often seen in contrastive learning, where embeddings converge to a single point and variance approaches

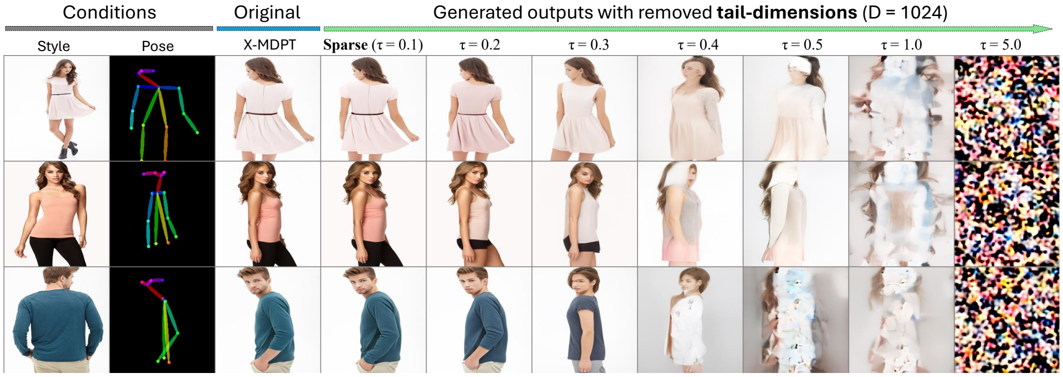
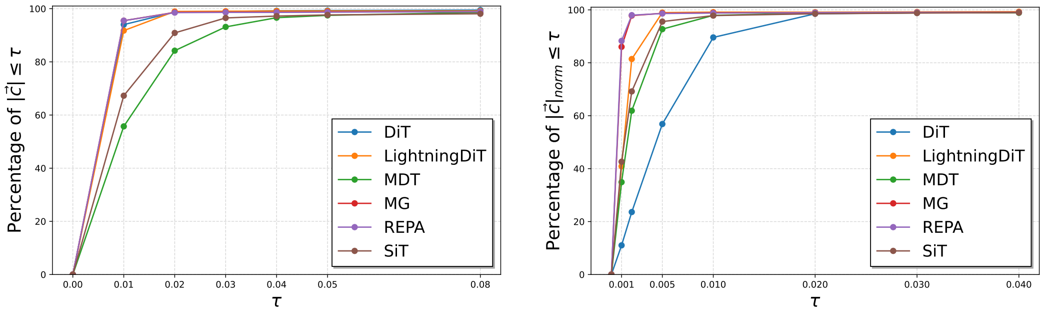


Figure 10: **Pose-conditioned person image generation.** *DeepFashion samples under different sparsification of $\vec{c}$.* τ is the magnitude threshold for pruning. Person-consistent images remain high quality even when 50–75% of dimensions are zeroed, as long as key head dimensions are preserved. Best viewed at 200% zoom; more samples appear in the Appendix.

zero (Wang & Isola, 2020; Zhang et al., 2022). Further, the dimension–threshold analysis in Fig. 11 reveals that with $\tau \approx 0.01$, a large fraction (50–90%) of dimensions retain low magnitude, saturating near $\tau = 0.02$, indicating that informative components are broadly distributed rather than concentrated in a few directions.

We hypothesize that diffusion transformers preserve this subtle but global structure because each denoising step predicts fine-grained Gaussian noise, providing a rich and stable training signal. Consequently, even when conditional embeddings lie within a narrow cone in feature space, their nuanced directional differences–amplified through adaptive layer normalization, the expressive Transformer backbone, and iterative refinement–remain sufficient to guide accurate class-conditional, pose-guided, and video-conditioned generation. *A deeper theoretical explanation of this robustness remains an open problem, calling for rigorous analysis in future work.*



(a) Dimensions below threshold τ (original scale). (b) Dimensions below threshold τ (normalized).

Figure 11: **Sparsity of conditional embeddings.** With $\tau = 0.01$, over 80% of components in the mean conditional vector fall below the threshold, revealing a highly sparse representation that starts to saturate near $\tau = 0.02$.

6.2 Hypothesizing the Structure of Conditional Embeddings

High Cosine Similarity. We hypothesize that the extreme cosine similarity among class embeddings arises from the dynamics of diffusion Transformer training (Fig. 12 top). Since the model conditions on embeddings across all timesteps t, it favors embeddings that provide a stable, robust signal for denoising, resulting in globally aligned embeddings:

$$\text{cosine}(c_y, c_{y'}) \approx 0.99 \quad \forall y \neq y'. \tag{3}$$

Despite this high similarity, semantic differences are encoded in a small subset of high-magnitude *head* dimensions,

$$c_y = c_{y,\text{head}} + c_{y,\text{tail}}, \quad \|c_{y,\text{head}}\| \gg \|c_{y,\text{tail}}\|, \tag{4}$$

are sufficient to modulate Adaptive LayerNorm parameters $\gamma(c_y), \beta(c_y)$:

$$\gamma(c_y) = W_\gamma c_y, \quad \beta(c_y) = W_\beta c_y. \tag{5}$$

These subtle differences are progressively amplified by the iterative diffusion process, enabling correct and high-quality class-conditional generation despite the high overall cosine similarity.

Observed sparsity in learned embeddings. Conditional embeddings are highly sparse: for $d = 1152$, only about 1–2% of dimensions reach large magnitudes ($\approx$ 5–8), while most remain near zero (10^{-3}–10^{-1}). This head–tail pattern indicates that semantic information resides in a small subspace, aligning with our pruning results. We quantify sparsity using the normalized participation ratio α_{norm}, which confirms that the effective dimensionality is far below d. As shown in Fig. 12 (bottom), monitoring this metric (nPR) while training the REPA B-2 model on ImageNet-1K for 200k steps shows a drop from about 90% early in training to under 6%, with the decline continuing–*revealing a natural sparsification dynamic in diffusion transformers.* Extended analyses appears in the Appendix.

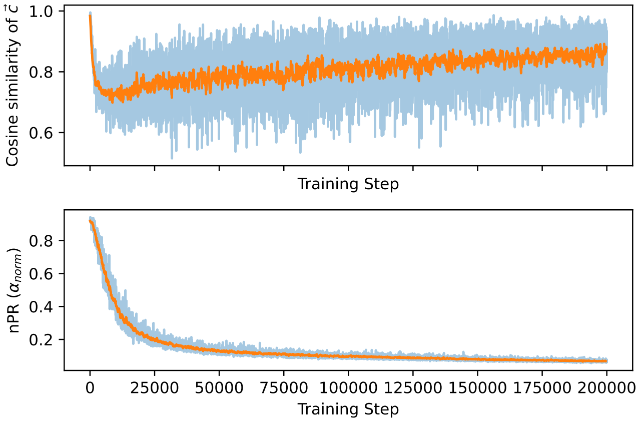


Figure 12: **Training dynamics of conditional embeddings.** Top: batchwise cosine similarity of $\vec{c}$ during training. Bottom: participation ratio (nPR) over training steps, showing progressive sparsification.

Pruning Improves Generation. We conjecture that pruning low-magnitude embedding dimensions acts as a form of noise suppression in diffusion Transformers. Let the conditional embedding be $c \in \mathbb{R}^d$ and decompose it as

$$c = c_{\text{head}} + c_{\text{tail}}, \quad \|c_{\text{head}}\| \gg \|c_{\text{tail}}\|. \tag{6}$$

In practice c is mapped to Adaptive LayerNorm parameters $\gamma(c), \beta(c)$ that modulate hidden states h:

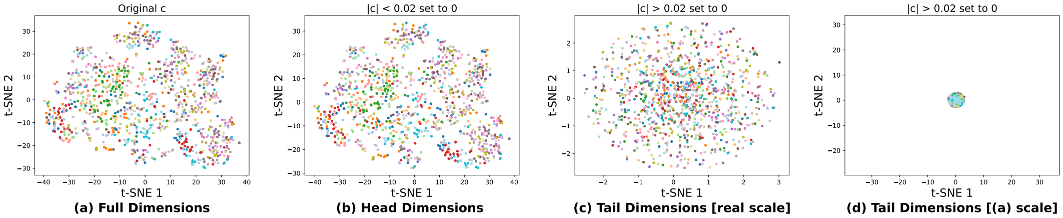
$$\text{AdaLN}(h \mid c) \;=\; \gamma(c) \odot \frac{h - \mu(h)}{\sigma(h)} + \beta(c). \tag{7}$$

Because semantic information concentrates in c_{head}, the tail c_{tail} contributes only weak, low-variance signals (as shown before in Fig. 9b). Supporting this view, Fig. 13 visualizes class embeddings: when only head dimensions are retained, class clusters remain well separated, whereas tail-only embeddings collapse into an entangled cloud. Retaining these noisy tail dimensions can perturb $\gamma(c), \beta(c)$ and inject interference into the denoising trajectory, particularly in later inference steps where precision is critical.

We empirically observe that pruning (zeroing out) c_{tail} at the initial step t_0 or at the final steps preserves generation quality, with late-step pruning yielding the strongest FID improvements. This supports the view that pruning suppresses interference and sharpens the semantic subspace. A more detailed analysis is provided in the Appendix.

7 Discussion

Our results reveal a semantic bottleneck in transformer-based diffusion models: conditional embeddings place most semantic content in a small set of high-magnitude dimensions, leaving the majority near zero and largely redundant. For class-conditional ImageNet generation, this effect is strongest, with only a few dominant dimensions and a very low normalized participation ratio (nPR). Continuous-condition tasks (e.g., pose-guided image or video-to-audio generation) show a milder form of this sparsity, exhibiting more high-magnitude



(a) Full Dimensions (b) Head Dimensions (c) Tail Dimensions [real scale] (d) Tail Dimensions [(a) scale]

Figure 13: **t-SNE of class embeddings by head vs. tail dimensions.** Keeping only head dimensions (b) preserves clear class clusters similar to the full embedding (a), while tail-only embeddings (c,d) collapse into entangled points, revealing weak semantic structure. Results are from SiT-XL on ImageNet-1K; similar trends appear in other models.

dimensions and higher nPR values. The consistency of this pattern across architectures and tasks points to a general property of how diffusion transformers encode conditioning signals.

Relation to Contrastive Learning Collapse The observed alignment of conditional embeddings bears resemblance to representation collapse in contrastive learning methods like SimCLR Chen et al. (2020), SimSiam (Chen & He, 2021), BYOL (Grill et al., 2020), VICReg (Bardes et al., 2022), and Barlow Twins (Zbontar et al., 2021). In those settings, collapse leads to trivial embeddings and degraded downstream performance unless variance-promoting regularizers or repulse components (negative samples) are used (Zhang et al., 2022). Interestingly, diffusion transformers avoid such pitfalls: despite extreme angular similarity, they maintain strong generation quality. We hypothesize that AdaLN amplifies high-magnitude dimensions sufficiently to preserve semantic distinctiveness during denoising, and that diffusion models' iterative refinement mitigates the impact of collapsed embeddings.

Why high cosine similarity occurs only in transformers. Based on additional experiments with U-Net models, we clarify that high cosine similarity arises primarily in transformers, and not in U-Nets when timestep embeddings are removed. However, similar redundancy emerges in U-Net diffusion models once timestep embeddings are included. This distinction appears tied to conditioning mechanisms: AdaLN in transformers promotes compression into dominant dimensions, whereas U-Nets use concatenation or cross-attention, preserving richer representations.

Relation to information bottleneck and AdaLN. The sparsity mirrors information bottleneck behavior (Tishby et al., 2000), where networks distill essential features. AdaLN's linear scaling amplifies a few dominant dimensions, rendering others redundant. **Implications and risks**. Compact embeddings may conflate unrelated semantics (cosine similarity $\neq$ semantic similarity). This could limit controllability in multi-conditional tasks or fine-grained editing.

Broader impact. Our observations of extreme similarity, sparse embeddings, and effective pruning in transformer-based diffusion models suggest that similar redundancy patterns could also appear in other generative frameworks, such as U-Net diffusion models (when timestep embeddings are included), GANs, or autoregressive models. This points to a potential principle of compact and efficient conditioning, which may inspire future work on lighter models and interpretable embeddings across tasks and modalities.

8 Conclusions

We have uncovered an interesting phenomenon in transformer-based diffusion models: extreme angular similarity and semantic sparsity in conditional embeddings. Our extensive analyses reveal that only a small subset of high-magnitude dimensions carry semantic information, while the majority of dimensions are redundant. Despite this, diffusion transformers maintain robust generation quality even when up to 66% of the conditional vector is pruned or masked. These findings suggest a fundamental overparameterization of conditional encoding and motivate rethinking conditioning mechanisms for efficiency and interpretability. Future architectures could benefit from compressed or hybrid conditioning strategies that maintain semantic fidelity while reducing computational overhead. Exploring these directions may lead to more controllable, efficient, and versatile generative models across vision, audio, and multimodal domains.

References

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub. 9

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020. 9

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 9

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23164–23173, 2023. 1, 2, 3, 14

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 2, 9

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

Zonglin Li, Chong You, Srinadh Bhojanapalli, Daliang Li, Ankit Singh Rawat, Sashank J. Reddi, Ke Ye, Felix Chern, Felix Yu, Ruiqi Guo, and Sanjiv Kumar. The lazy neuron phenomenon: On emergence of activation sparsity in transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TJ2nxciYCk-. 2

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017. 2

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024. 1, 2, 3

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. 1, 2, 3, 14, 15

Trung X. Pham, Kang Zhang, and Chang D. Yoo. Cross-view masked diffusion transformers for person image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=jEoIkNkqyc. 1, 2, 3, 4, 6, 27

Trung X. Pham, Tri Ton, and Chang D. Yoo. MDSGen: Fast and efficient masked diffusion temporal-aware transformers for open-domain sound generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yFEqYwgttJ. 1, 2, 3

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022. 2, 3

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. 2

Peng Sun, Yi Jiang, and Tao Lin. Unified continuous generative models. *arXiv preprint arXiv:2505.07447*, 2025. 2

Zhicong Tang, Jianmin Bao, Dong Chen, and Baining Guo. Diffusion models without classifier-free guidance. *arXiv preprint arXiv:2502.12154*, 2025. 1, 2, 3, 15, 18, 27

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 2, 9

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, pp. 6558, 2019. 2

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020. 7

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025. 1, 2, 3, 12, 14, 18, 27

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. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DJSZGGZYVi. 1, 2, 3, 15, 27

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *International Conference on Machine Learning*, 2021. 2, 9

Chaoning Zhang, Kang Zhang, Chenshuang Zhang, Trung X. Pham, Chang D. Yoo, and In So Kweon. How does simsiam avoid collapse without negative samples? a unified understanding with self-supervised contrastive learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=bwq6O4Cwdl. 7, 9

A Appendix

In this appendix, we provide the details of experimental setups as well as more comprehensive analysis results for various approaches.

A.1 Setup Details

We generate 5,000 samples from the public checkpoints of each method (5 samples per ImageNet class) and evaluate FID and IS using the LightningDiT (Yao et al., 2025) evaluation code. Inference follows the default hyperparameters and sampling steps specified by each model, using XL-size variants when available. For continuous tasks (X-MDPT and MDSGen with AdaLN), we adopt their largest released models (L-size and B-size, respectively). During inference, we modify only the conditional vector $\vec{c}$, keeping all other components unchanged.

A.2 Explanation Hypothesis (Extended)

Pruning Improves Generation. We retain the decomposition used throughout the paper:

$$c = c_{\text{head}} + c_{\text{tail}}, \quad \|c_{\text{head}}\| \gg \|c_{\text{tail}}\|.$$

Here c_{head} denotes the high-variance, semantically informative dimensions, while c_{tail} corresponds to low-magnitude, low-variance dimensions.

Conditioning is implemented through Adaptive Layer Normalization (AdaLN). For a hidden activation $h \in \mathbb{R}^d$:

$$\text{AdaLN}(h \mid c) = \gamma(c) \odot \frac{h - \mu(h)}{\sigma(h)} + \beta(c),$$

with linear projections

$$\gamma(c) = W_\gamma c, \qquad \beta(c) = W_\beta c.$$

Linearity implies that

$$\gamma(c) = \gamma(c_{\text{head}}) + \gamma(c_{\text{tail}}), \qquad \beta(c) = \beta(c_{\text{head}}) + \beta(c_{\text{tail}}).$$

Empirically, $\text{Var}[\gamma(c_{\text{tail}})]$ and $\text{Var}[\beta(c_{\text{tail}})]$ are negligible compared to their head counterparts. However, we hypothesize that these weak terms can propagate as noise through the denoising trajectory, with a potentially larger effect in later inference steps ($t \to 0$), where precision is critical.

Define a pruning operator $\mathcal{P}(\cdot)$ that zeros out tail dimensions:

$$c' = \mathcal{P}(c) = c_{\text{head}}.$$

Pruning can be applied either at the initial step t_0 or during the final few steps of inference. While early-step pruning reduces redundancy early, we empirically observe that late-step pruning consistently yields stronger improvements in FID, supporting the hypothesis that late-step pruning suppresses residual noise and sharpens semantic guidance:

$$\text{AdaLN}(h \mid c') = \gamma(c_{\text{head}}) \odot \frac{h - \mu(h)}{\sigma(h)} + \beta(c_{\text{head}}).$$

Thus, pruning acts as an effective noise filter: removing weak tail dimensions reduces interference while focusing conditioning on dominant semantic directions, explaining why pruning preserves or can even improve generative quality.

High Cosine Similarity. We empirically observe that the cosine similarity between class embeddings remains extremely high (> 0.99) across nearly all timesteps of denoising. We hypothesize that this is a consequence of dynamic training in diffusion Transformers: conditioning is applied across all timesteps t, and the network learns to maintain a stable, robust signal. This encourages embeddings to align along similar directions, while semantic distinctions are preserved in a small subspace of head dimensions:

$$c_y = c_{y,\text{head}} + c_{y,\text{tail}}, \quad \|c_{y,\text{head}}\| \gg \|c_{y,\text{tail}}\|.$$

Even with globally aligned embeddings, the high-magnitude head dimensions provide sufficient directional cues to modulate Adaptive LayerNorm parameters:

$$\gamma(c_y) = W_\gamma c_y, \quad \beta(c_y) = W_\beta c_y.$$

These small differences are progressively amplified by the iterative denoising process.

Thus, while embeddings appear nearly parallel in the full space, the effective semantic subspace defined by the head dimensions ensures that generation remains accurate and high-quality. This also explains why pruning tail dimensions, which contain low-magnitude, redundant signals, does not harm generation and can sometimes improve quality.

A.3 Analysis of Embedding Sparsity

Empirical observation. In the pretrained diffusion Transformer embeddings we analyze ($d = 1152$), only a small subset of dimensions—about 1% to 2%—exhibit large absolute values (typical magnitude $\sim$ 5–8), while the rest remain near-zero (typical magnitude $\sim 10^{-3}$–10^{-1}). We refer to the large-magnitude coordinates as the *head* and the rest as the *tail*.

Metrics. To quantify sparsity and effective dimensionality, we use the following statistics:

- **Sparsity ratio** at threshold τ:

  $$s(\tau) = \frac{1}{d}\#\{i : |c_i| > \tau\}.$$

  With τ set to a small constant (e.g., 0.5) this yields $s \approx 0.01$–0.02 empirically.

- **Participation ratio** (PR) on absolute magnitudes $v_i = |c_i|$ (a measure of effective dimensions):

  $$\alpha = \mathrm{PR}(v) = \frac{\left(\sum_{i=1}^{d} v_i\right)^2}{\sum_{i=1}^{d} v_i^2}, \ \alpha_{\text{normalized}} = \frac{1}{d} \times \mathrm{PR}(v).$$

  PR gives an estimate of how many coordinates carry most of the total magnitude; we find PR $\ll d$ (order tens).

This normalization maps the range of values to $\alpha_{\text{normalized}} \in (0, 1]$, nPR ($\alpha_{\text{normalized}}$) $= 1$ when all d coordinates contribute equally. nPR $\approx k/d$ when effectively only k coordinates carry the magnitude.

Interpretation hypotheses. We offer several plausible, non-exclusive explanations for this sparse phenomenon:

1. **Projection and scale effects.** The learned linear projections W_γ, W_β (and any subsequent layers) can amplify a few coordinates of c if their corresponding projection weights are large, producing a few dominant coordinates in the final modulation parameters.
2. **Stable conditioning across timesteps.** Because conditioning is applied across many timesteps, the optimizer favors a stable, low-dimensional conditioning signal to avoid disturbing denoising dynamics; encoding semantics in a few robust axes avoids noisy, volatile conditioning.
3. **Implicit sparsity from optimization/regularization.** Weight decay, initialization, and training dynamics may implicitly encourage small-magnitude coordinates; only the coordinates providing robust semantic signal are driven to large magnitudes.

Consequences for AdaLN modulation. Since AdaLN uses $\gamma(c) = W_\gamma c$ and $\beta(c) = W_\beta c$, large entries in c dominate the modulation:

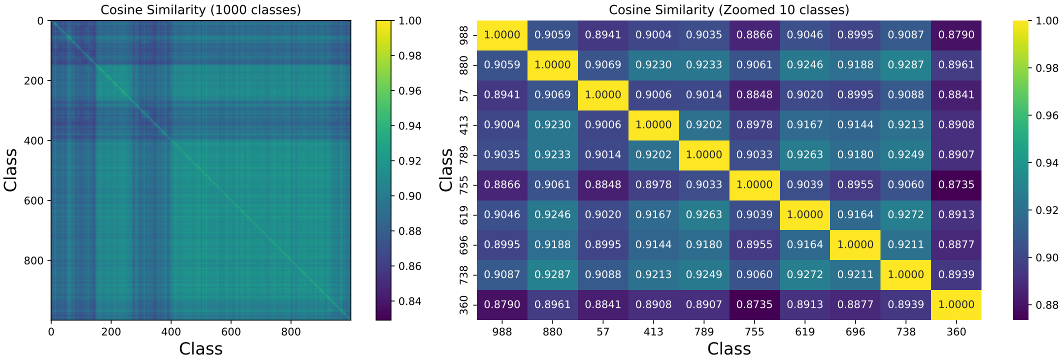
$$\gamma(c) \approx W_\gamma c_{\text{head}}, \qquad \beta(c) \approx W_\beta c_{\text{head}}.$$

Thus, the denoising network effectively receives conditioning from a low-dimensional subspace, explaining why zeroing most coordinates (sparsification) has a small empirical impact and why pruning can even improve performance by removing weak, noisy contributions.

A.4 More Visualizations of Other Methods

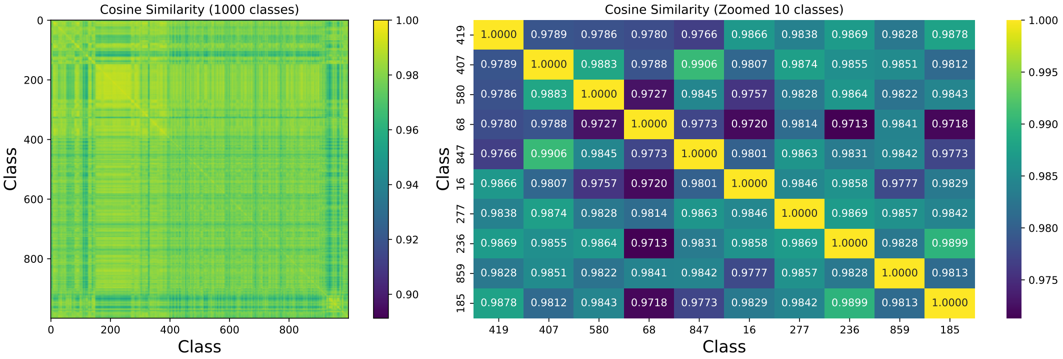
A.4.1 Cosine Similarity with Pairwise Analysis.

For completeness, we present full pairwise cosine similarity matrices for all six state-of-the-art diffusion transformer models evaluated on ImageNet. Each matrix reports the cosine similarity between conditional embeddings for every pair of ImageNet-1K classes, offering a comprehensive view of how uniformly aligned these vectors are across the label space. The results reinforce the main paper's findings: near-uniform similarity is pervasive across models and classes, with the sole exception of DiT, whose lowest pairwise similarity is about 88%. Notably, DiT also delivers weaker generative performance (higher FID) compared to the other models, further distinguishing it from the rest of the evaluated methods.



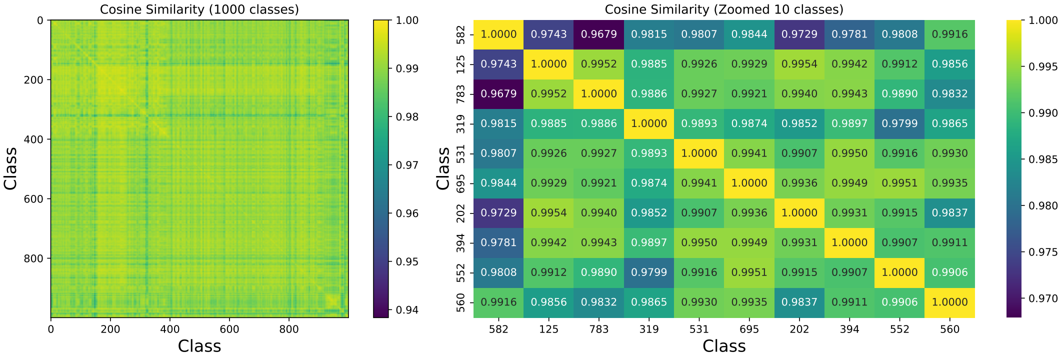
	988	880	57	413	789	755	619	696	738	360
988	1.0000	0.9059	0.8941	0.9004	0.9035	0.8866	0.9046	0.8995	0.9087	0.8790
880	0.9059	1.0000	0.9069	0.9230	0.9233	0.9061	0.9246	0.9188	0.9287	0.8961
57	0.8941	0.9069	1.0000	0.9006	0.9014	0.8848	0.9020	0.8995	0.9088	0.8841
413	0.9004	0.9230	0.9006	1.0000	0.9202	0.8978	0.9167	0.9144	0.9213	0.8908
789	0.9035	0.9233	0.9014	0.9202	1.0000	0.9033	0.9263	0.9180	0.9249	0.8907
755	0.8866	0.9061	0.8848	0.8978	0.9033	1.0000	0.9039	0.8955	0.9060	0.8735
619	0.9046	0.9246	0.9020	0.9167	0.9263	0.9039	1.0000	0.9164	0.9272	0.8913
696	0.8995	0.9188	0.8995	0.9144	0.9180	0.8955	0.9164	1.0000	0.9211	0.8877
738	0.9087	0.9287	0.9088	0.9213	0.9249	0.9060	0.9272	0.9211	1.0000	0.8939
360	0.8790	0.8961	0.8841	0.8908	0.8907	0.8735	0.8913	0.8877	0.8939	1.0000

Figure 14: Cosine similarity of conditional vectors $\vec{c} = y + t$ across 1000 ImageNet classes using DiT-XL (Peebles & Xie, 2023).



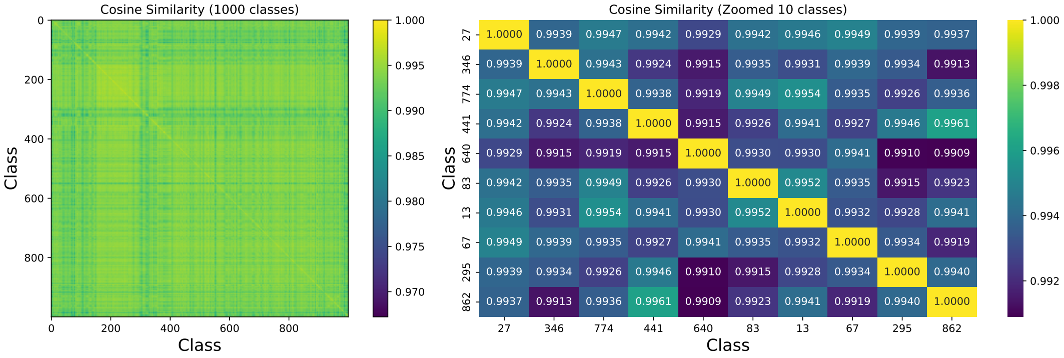
	419	407	580	68	847	16	277	236	859	185
419	1.0000	0.9789	0.9786	0.9780	0.9766	0.9866	0.9838	0.9869	0.9828	0.9878
407	0.9789	1.0000	0.9883	0.9788	0.9906	0.9807	0.9874	0.9855	0.9851	0.9812
580	0.9786	0.9883	1.0000	0.9727	0.9845	0.9757	0.9828	0.9864	0.9822	0.9843
68	0.9780	0.9788	0.9727	1.0000	0.9773	0.9720	0.9814	0.9713	0.9841	0.9718
847	0.9766	0.9906	0.9845	0.9773	1.0000	0.9801	0.9863	0.9831	0.9842	0.9773
16	0.9866	0.9807	0.9757	0.9720	0.9801	1.0000	0.9846	0.9858	0.9777	0.9829
277	0.9838	0.9874	0.9828	0.9814	0.9863	0.9846	1.0000	0.9869	0.9857	0.9842
236	0.9869	0.9855	0.9864	0.9713	0.9831	0.9858	0.9869	1.0000	0.9828	0.9899
859	0.9828	0.9851	0.9822	0.9841	0.9842	0.9777	0.9857	0.9828	1.0000	0.9813
185	0.9878	0.9812	0.9843	0.9718	0.9773	0.9829	0.9842	0.9899	0.9813	1.0000

Figure 15: Cosine similarity of conditional vectors $\vec{c} = y + t$ across 1000 ImageNet classes using LightningDiT-XL (Yao et al., 2025).



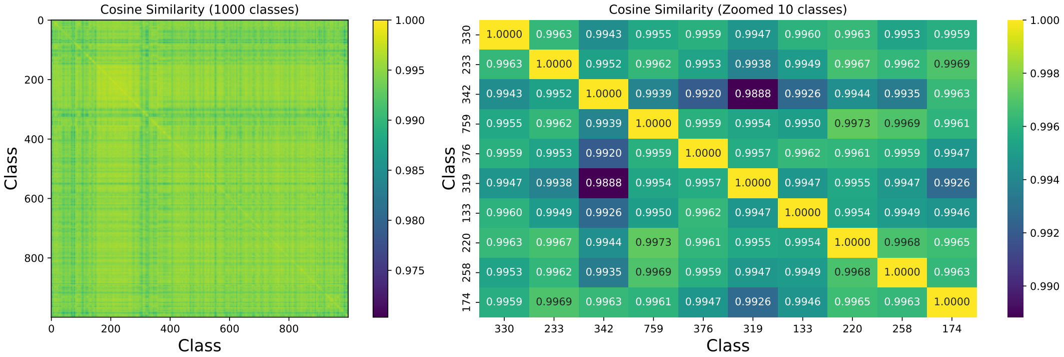
	582	125	783	319	531	695	202	394	552	560
582	1.0000	0.9743	0.9679	0.9815	0.9807	0.9844	0.9729	0.9781	0.9808	0.9916
125	0.9743	1.0000	0.9952	0.9885	0.9926	0.9929	0.9954	0.9942	0.9912	0.9856
783	0.9679	0.9952	1.0000	0.9886	0.9927	0.9921	0.9940	0.9943	0.9890	0.9832
319	0.9815	0.9885	0.9886	1.0000	0.9893	0.9874	0.9852	0.9897	0.9799	0.9865
531	0.9807	0.9926	0.9927	0.9893	1.0000	0.9941	0.9907	0.9950	0.9916	0.9930
695	0.9844	0.9929	0.9921	0.9874	0.9941	1.0000	0.9936	0.9949	0.9951	0.9935
202	0.9729	0.9954	0.9940	0.9852	0.9907	0.9936	1.0000	0.9931	0.9915	0.9837
394	0.9781	0.9942	0.9943	0.9897	0.9950	0.9949	0.9931	1.0000	0.9907	0.9911
552	0.9808	0.9912	0.9890	0.9799	0.9916	0.9951	0.9915	0.9907	1.0000	0.9906
560	0.9916	0.9856	0.9832	0.9865	0.9930	0.9935	0.9837	0.9911	0.9906	1.0000

Figure 16: Cosine similarity of conditional vectors $\vec{c} = y + t$ across 1000 ImageNet classes using MDT-XL (Gao et al., 2023).



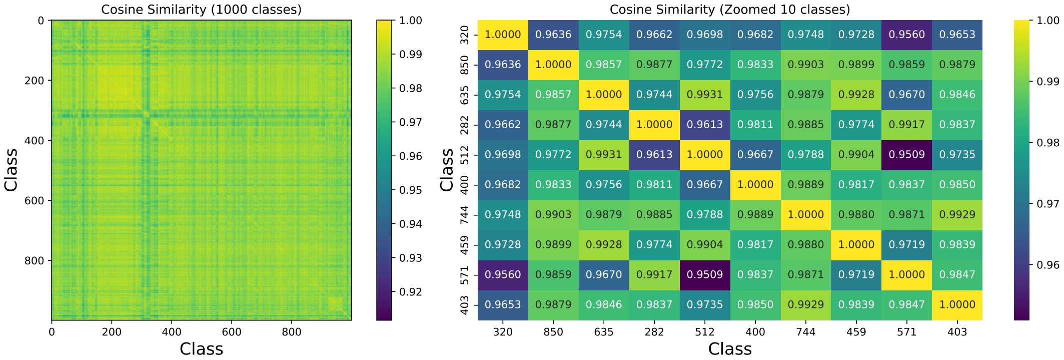
	27	346	774	441	640	83	13	67	295	862
27	1.0000	0.9939	0.9947	0.9942	0.9929	0.9942	0.9946	0.9949	0.9939	0.9937
346	0.9939	1.0000	0.9943	0.9924	0.9915	0.9935	0.9931	0.9939	0.9934	0.9913
774	0.9947	0.9943	1.0000	0.9938	0.9919	0.9949	0.9954	0.9935	0.9926	0.9936
441	0.9942	0.9924	0.9938	1.0000	0.9915	0.9926	0.9941	0.9927	0.9946	0.9961
640	0.9929	0.9915	0.9919	0.9915	1.0000	0.9930	0.9930	0.9941	0.9910	0.9909
83	0.9942	0.9935	0.9949	0.9926	0.9930	1.0000	0.9952	0.9935	0.9915	0.9923
13	0.9946	0.9931	0.9954	0.9941	0.9930	0.9952	1.0000	0.9932	0.9928	0.9941
67	0.9949	0.9939	0.9935	0.9927	0.9941	0.9935	0.9932	1.0000	0.9934	0.9919
295	0.9939	0.9934	0.9926	0.9946	0.9910	0.9915	0.9928	0.9934	1.0000	0.9940
862	0.9937	0.9913	0.9936	0.9961	0.9909	0.9923	0.9941	0.9919	0.9940	1.0000

Figure 17: Cosine similarity of conditional vectors $\vec{c} = y + t$ across 1000 ImageNet classes using MG-XL (Tang et al., 2025).



	330	233	342	759	376	319	133	220	258	174
330	1.0000	0.9963	0.9943	0.9955	0.9959	0.9947	0.9960	0.9963	0.9953	0.9959
233	0.9963	1.0000	0.9952	0.9962	0.9953	0.9938	0.9949	0.9967	0.9962	0.9969
342	0.9943	0.9952	1.0000	0.9939	0.9920	0.9888	0.9926	0.9944	0.9935	0.9963
759	0.9955	0.9962	0.9939	1.0000	0.9959	0.9954	0.9950	0.9973	0.9969	0.9961
376	0.9959	0.9953	0.9920	0.9959	1.0000	0.9957	0.9962	0.9961	0.9959	0.9947
319	0.9947	0.9938	0.9888	0.9954	0.9957	1.0000	0.9947	0.9955	0.9947	0.9926
133	0.9960	0.9949	0.9926	0.9950	0.9962	0.9947	1.0000	0.9954	0.9949	0.9946
220	0.9963	0.9967	0.9944	0.9973	0.9961	0.9955	0.9954	1.0000	0.9968	0.9965
258	0.9953	0.9962	0.9935	0.9969	0.9959	0.9947	0.9949	0.9968	1.0000	0.9963
174	0.9959	0.9969	0.9963	0.9961	0.9947	0.9926	0.9946	0.9965	0.9963	1.0000

Figure 18: Cosine similarity of conditional vectors $\vec{c} = y + t$ across 1000 ImageNet classes using REPA-XL (Yu et al., 2025).



	320	850	635	282	512	400	744	459	571	403
320	1.0000	0.9636	0.9754	0.9662	0.9698	0.9682	0.9748	0.9728	0.9560	0.9653
850	0.9636	1.0000	0.9857	0.9877	0.9772	0.9833	0.9903	0.9899	0.9859	0.9879
635	0.9754	0.9857	1.0000	0.9744	0.9931	0.9756	0.9879	0.9928	0.9670	0.9846
282	0.9662	0.9877	0.9744	1.0000	0.9613	0.9811	0.9885	0.9774	0.9917	0.9837
512	0.9698	0.9772	0.9931	0.9613	1.0000	0.9667	0.9788	0.9904	0.9509	0.9735
400	0.9682	0.9833	0.9756	0.9811	0.9667	1.0000	0.9889	0.9817	0.9837	0.9850
744	0.9748	0.9903	0.9879	0.9885	0.9788	0.9889	1.0000	0.9880	0.9871	0.9929
459	0.9728	0.9899	0.9928	0.9774	0.9904	0.9817	0.9880	1.0000	0.9719	0.9839
571	0.9560	0.9859	0.9670	0.9917	0.9509	0.9837	0.9871	0.9719	1.0000	0.9847
403	0.9653	0.9879	0.9846	0.9837	0.9735	0.9850	0.9929	0.9839	0.9847	1.0000

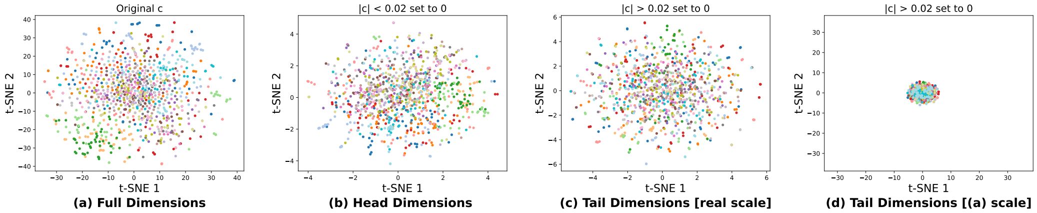
Figure 19: Cosine similarity of conditional vectors $\vec{c} = y + t$ across 1000 ImageNet classes using SiT-XL (Peebles & Xie, 2023).

A.4.2 t-SNE Distribution Analysis.

To further examine the role of head and tail dimensions in conditional embeddings, we provide t-SNE visualizations for all evaluated methods under targeted perturbations. Specifically, we manipulate either the high-magnitude (head) or low-magnitude (tail) dimensions of the embeddings and observe how these changes affect the overall distribution of class representations for all 1,000 ImageNet classes.

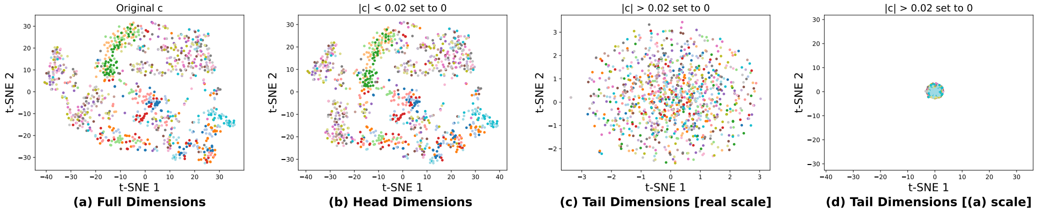
These visualizations illustrate that removing or altering head dimensions strongly disrupts the separability of class clusters, while perturbing tail dimensions has minimal impact, highlighting the concentration of semantic information in a small subset of dimensions.

Results for each method are shown in Fig. 20 – Fig. 25, providing a comparative view of how different architectures encode and distribute semantic information in their conditional embeddings.



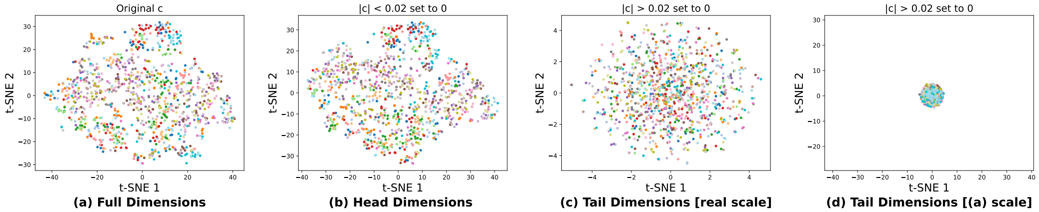
(a) Full Dimensions (b) Head Dimensions (c) Tail Dimensions [real scale] (d) Tail Dimensions [(a) scale]

Figure 20: **(DiT) t-SNE of class embeddings by head vs. tail dimensions.** Keeping only head dimensions (b) preserves clear class clusters similar to the full embedding (a), while tail-only embeddings (c,d) collapse into entangled points, revealing weak semantic structure. Results are from DiT-XL on ImageNet-1K; similar trends appear in other models.



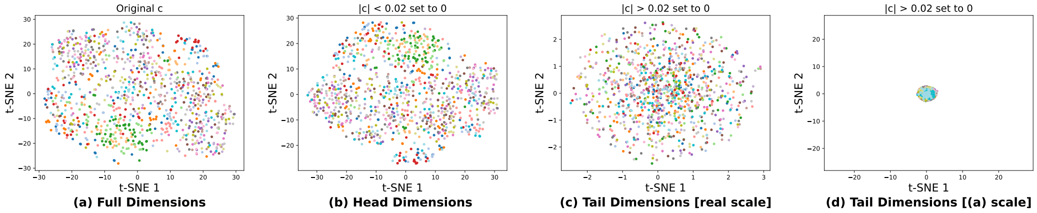
(a) Full Dimensions (b) Head Dimensions (c) Tail Dimensions [real scale] (d) Tail Dimensions [(a) scale]

Figure 21: **(LightningDiT) t-SNE of class embeddings by head vs. tail dimensions.** Keeping only head dimensions (b) preserves clear class clusters similar to the full embedding (a), while tail-only embeddings (c,d) collapse into entangled points, revealing weak semantic structure. Results are from LightningDiT-XL on ImageNet-1K; similar trends appear in other models.



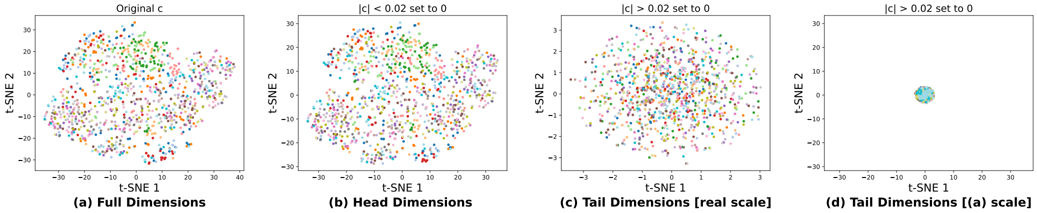
(a) Full Dimensions (b) Head Dimensions (c) Tail Dimensions [real scale] (d) Tail Dimensions [(a) scale]

Figure 22: **(MDT) t-SNE of class embeddings by head vs. tail dimensions.** Keeping only head dimensions (b) preserves clear class clusters similar to the full embedding (a), while tail-only embeddings (c,d) collapse into entangled points, revealing weak semantic structure. Results are from MDT-XL on ImageNet-1K; similar trends appear in other models.



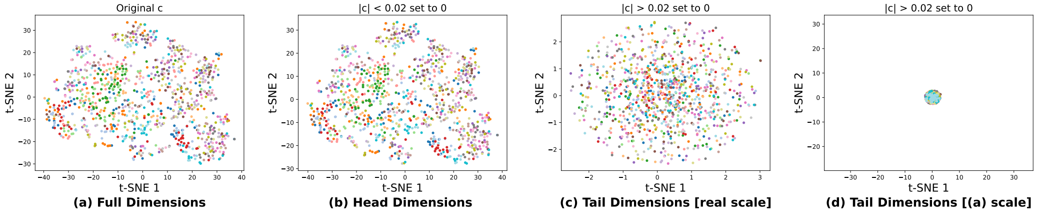
(a) Full Dimensions (b) Head Dimensions (c) Tail Dimensions [real scale] (d) Tail Dimensions [(a) scale]

Figure 23: **(MG) t-SNE of class embeddings by head vs. tail dimensions.** Keeping only head dimensions (b) preserves clear class clusters similar to the full embedding (a), while tail-only embeddings (c,d) collapse into entangled points, revealing weak semantic structure. Results are from MG-XL on ImageNet-1K; similar trends appear in other models.



(a) Full Dimensions (b) Head Dimensions (c) Tail Dimensions [real scale] (d) Tail Dimensions [(a) scale]

Figure 24: **(REPA) t-SNE of class embeddings by head vs. tail dimensions.** Keeping only head dimensions (b) preserves clear class clusters similar to the full embedding (a), while tail-only embeddings (c,d) collapse into entangled points, revealing weak semantic structure. Results are from REPA-XL on ImageNet-1K; similar trends appear in other models.



(a) Full Dimensions (b) Head Dimensions (c) Tail Dimensions [real scale] (d) Tail Dimensions [(a) scale]

Figure 25: **(SiT) t-SNE of class embeddings by head vs. tail dimensions.** Keeping only head dimensions (b) preserves clear class clusters similar to the full embedding (a), while tail-only embeddings (c,d) collapse into entangled points, revealing weak semantic structure. Results are from SiT-XL on ImageNet-1K; similar trends appear in other models.

A.4.3 Additional Baselines with Sparse Conditioning

We extend the evaluation from the main paper to two additional strong baselines: LightningDiT (Yao et al., 2025) and MG (Tang et al., 2025), following the same experimental protocol. As shown in Tab. 3, pruning low-magnitude dimensions in the conditional vector consistently improves both FID and CLIP scores. These results reinforce our main finding that dense conditional embeddings contain noisy, low-utility dimensions and that sparsification can yield more efficient and effective generative models.

Table 3: **More baselines.** Performance and semantic metrics under sparsification. t_i: prune every step, t_0: prune only at start, $t_{n-k,n}$: prune during last k steps.

Prune	**Threshold τ**	**# Removed Dims**	**FID ↓**	**IS ↑**	**CLIP↑**
	Baseline MG (Tang et al., 2025)	0/1152 (0%)	7.2478	174.5151	**30.199**
Tail	$\tau = 0.01$ (t_i)	448/1152 (38.94%)	7.2791	170.55	30.140
	$\tau = 0.01$ (t_0)	448/1152 (38.94%)	**7.2466**	**174.5537**	**30.199**
	$\tau = 0.01$ ($t_{n-k,n}$)	448/1152 (38.94%)	**7.2455**	174.3103	**30.198**
.	Baseline LightningDiT (Yao et al., 2025)	0/1152 (0%)	7.0802	169.8574	30.720
Tail	$\tau = 0.01$ (t_i)	448/1152 (38.94%)	**7.0130**	166.0569	30.7045
	$\tau = 0.01$ (t_0)	448/1152 (38.94%)	**7.0712**	169.9164	**30.729**
	$\tau = 0.01$ ($t_{n-k,n}$)	448/1152 (38.94%)	**7.0745**	**169.9236**	**30.729**

A.4.4 Variance Distribution Analysis.

We analyze the per-dimension variance of the conditional embeddings by first computing the mean vector for each method and then measuring the variance across classes for each dimension. As expected, high-magnitude dimensions (head dimensions) exhibit substantially higher variance than the low-magnitude (tail) dimensions, reinforcing the observation that semantic information is concentrated in the head.

An exception is DiT, where the conditional vectors have smaller absolute values (maximum around 0.8, compared to 4–8 for other models), resulting in a different variance pattern. These results, visualized in Fig. 26 to Fig. 31, provide further evidence of the head–tail structure and its connection to semantic encoding in diffusion transformer embeddings.

For continuous-condition tasks such as pose-guided person image generation and video-guided audio generation, the learned embeddings are noticeably less sparse, consistent with the higher participation-ratio scores reported in Tab. 1 of the main paper. Detailed variance and mean analyses for these tasks are provided in Fig. 33 and Fig. 32.

A.5 Additional Qualitative Results

We present an extended set of qualitative results for both class-conditional image generation on ImageNet and pose-guided person image synthesis. These visualizations highlight the impact of pruning low-magnitude dimensions in the conditional embedding vector.

Across a wide range of samples, we observe that removing these tail dimensions often preserves the generation quality and, in some cases, even enhances visual fidelity or sharpness. This supports our main finding that semantic information is concentrated in a small subset of head dimensions, while the majority of the embedding space is redundant.

Representative examples are provided in Fig. 34 through Fig. 41, demonstrating consistent trends across different models, classes, and poses.

A.6 Use of Large Language Models

Large Language Models (LLMs) were used solely as a writing-assistance tool to polish grammar and improve sentence clarity. All research ideas, experimental design, analyses, and results were conceived and executed entirely by the authors. The LLM did not contribute to research ideation, data analysis, or the generation of any scientific content.

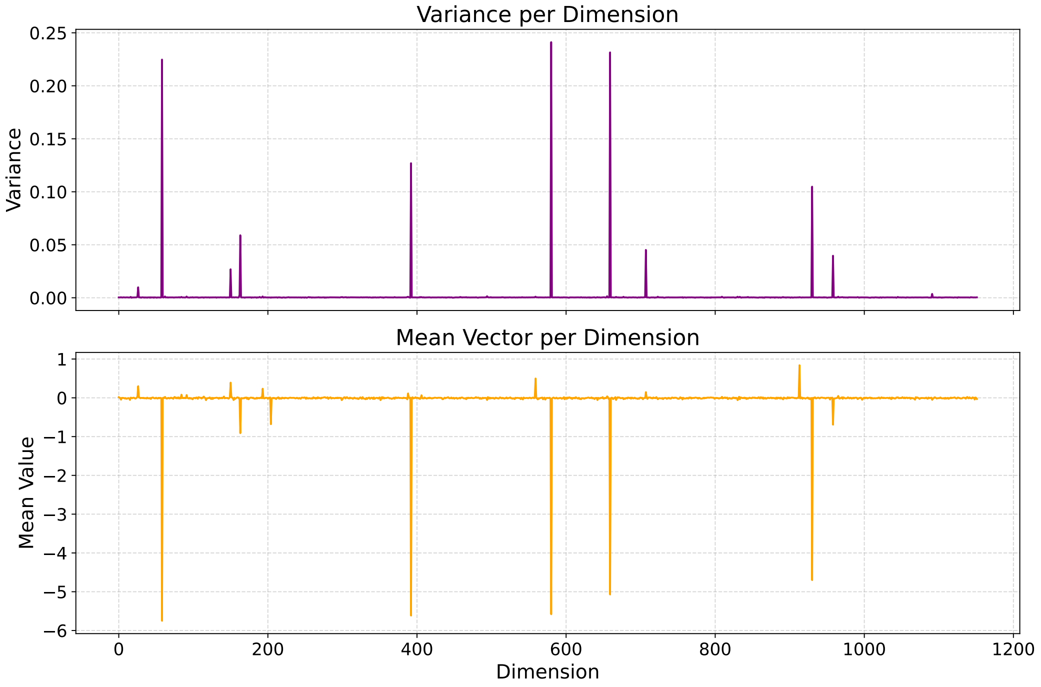


Figure 26: Variance per dimension of the conditional vector learned by MDT-XL.

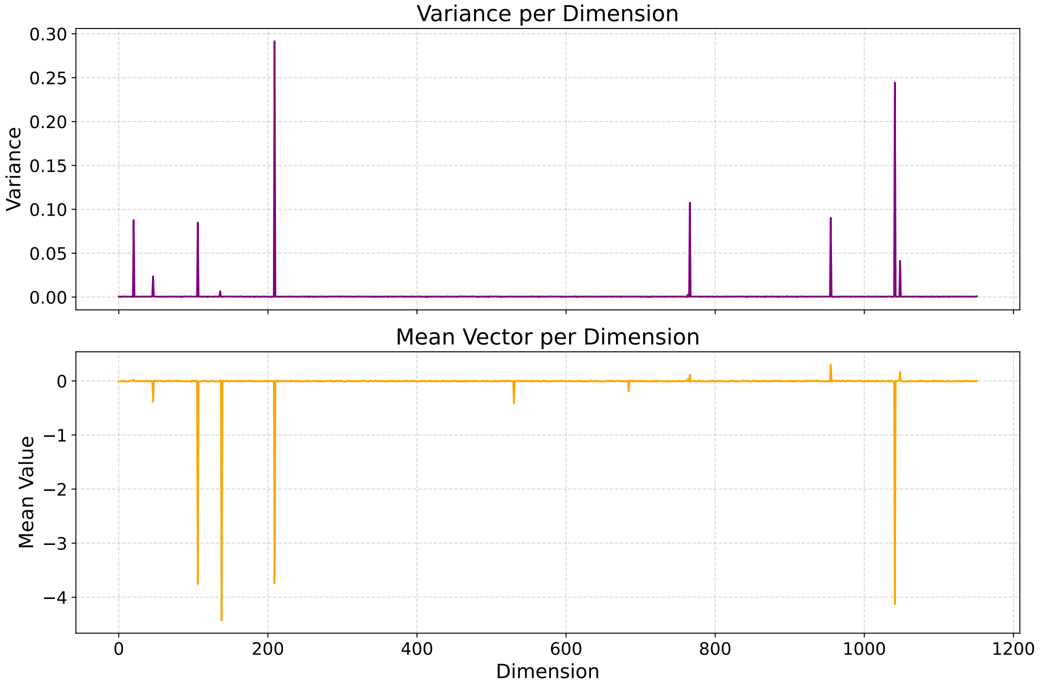


Figure 27: Variance per dimension of the conditional vector learned by LightningDiT-XL.

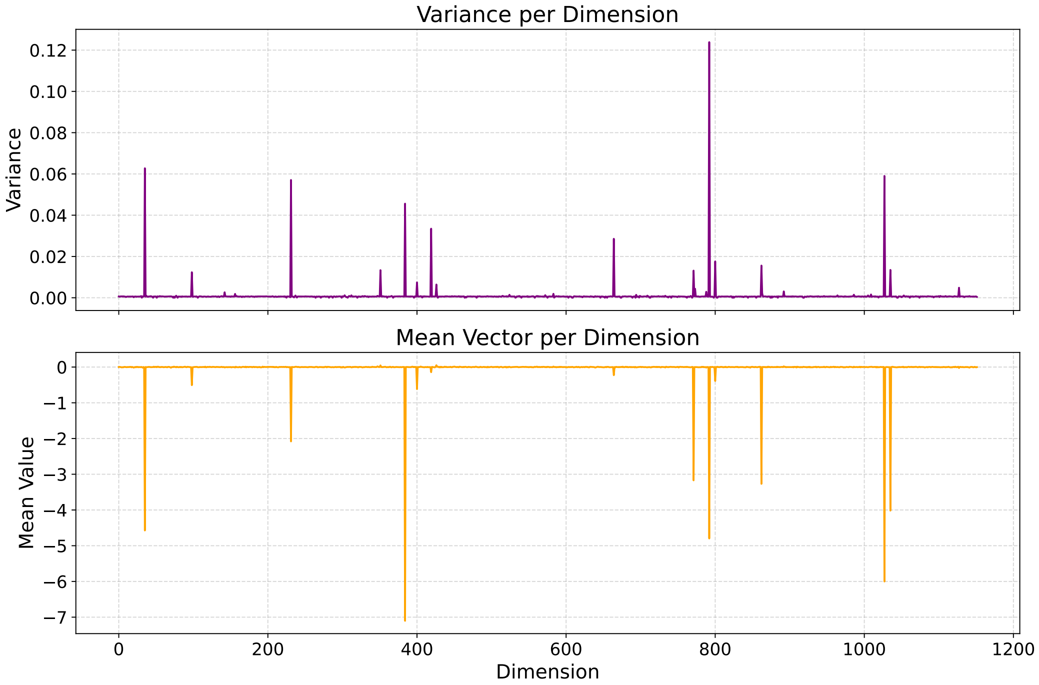


Figure 28: Variance per dimension of the conditional vector learned by MG-XL.

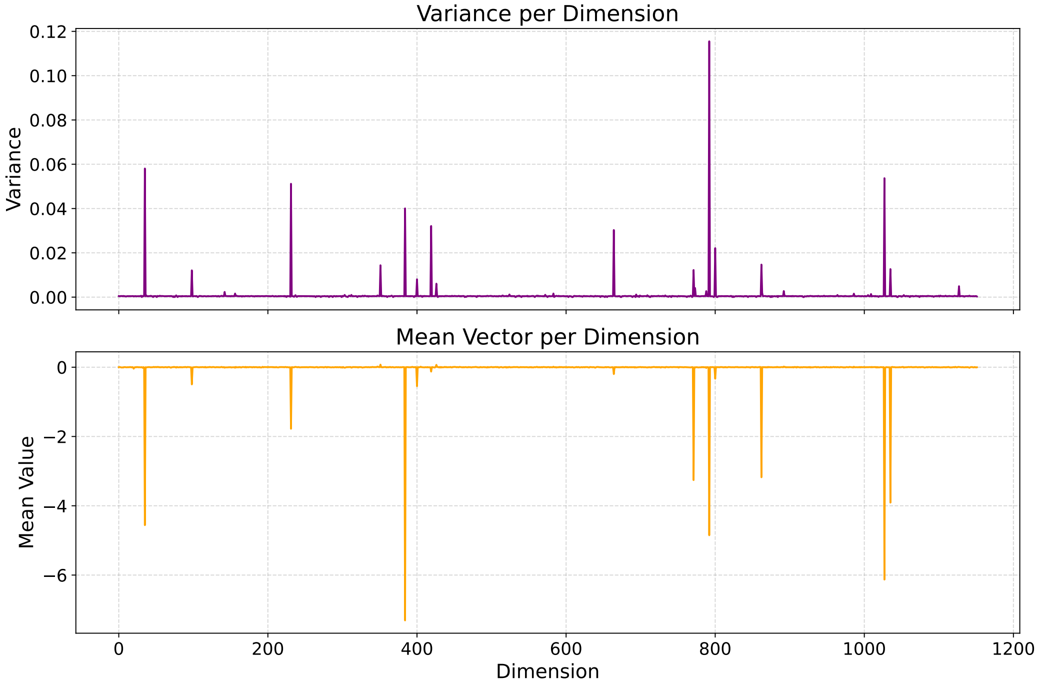


Figure 29: Variance per dimension of the conditional vector learned by REPA-XL.

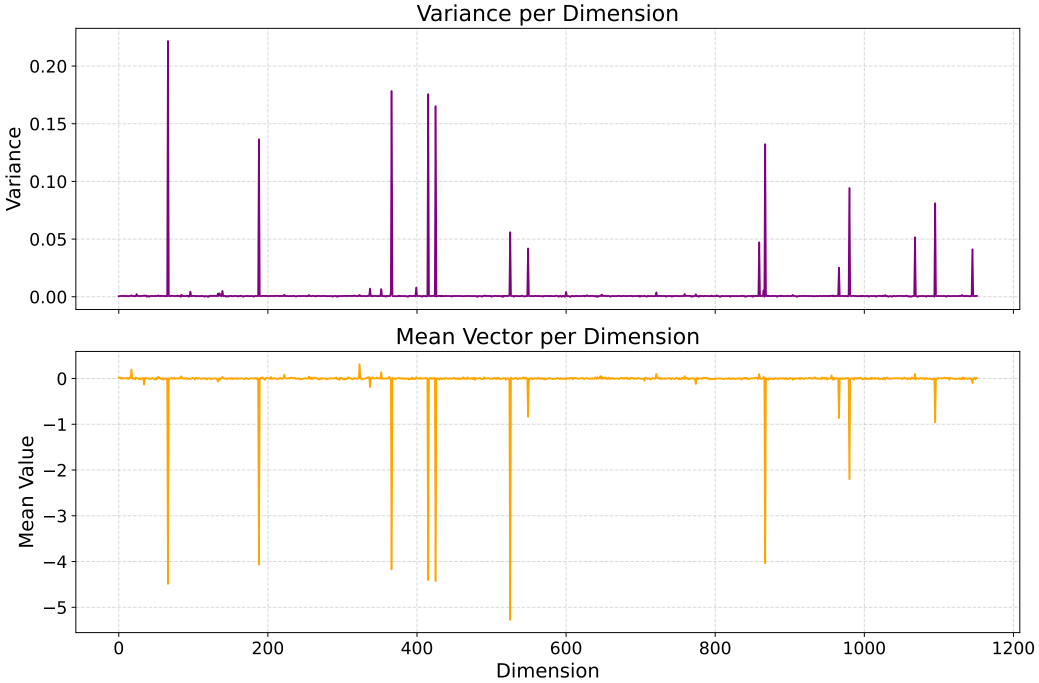


Figure 30: Variance per dimension of the conditional vector learned by SiT-XL.

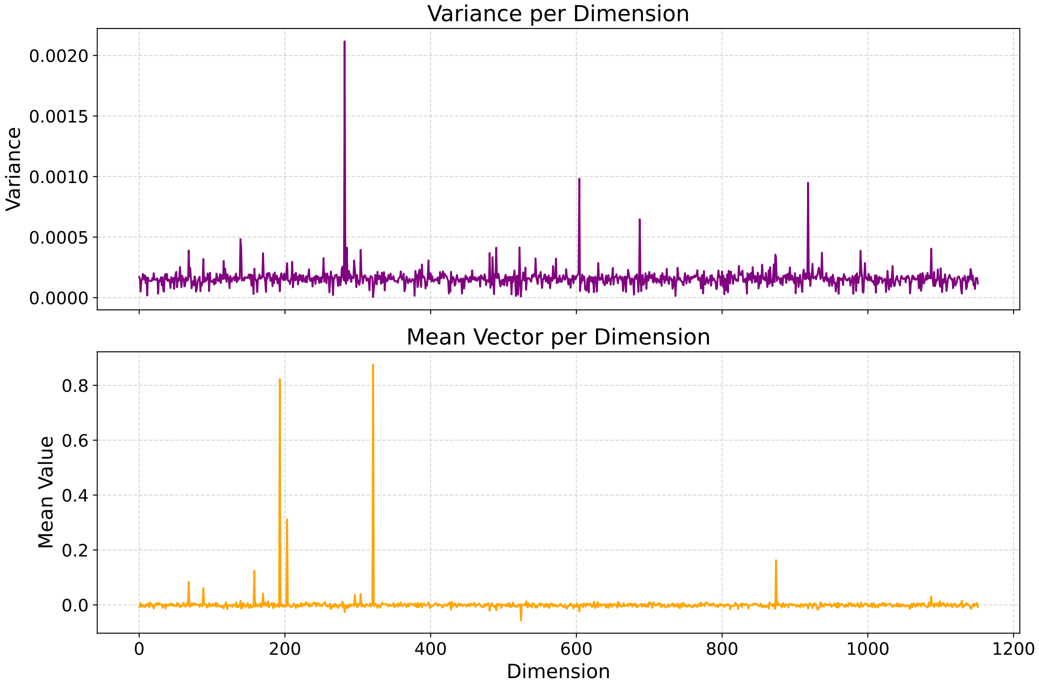


Figure 31: Variance per dimension of the conditional vector learned by DiT-XL.

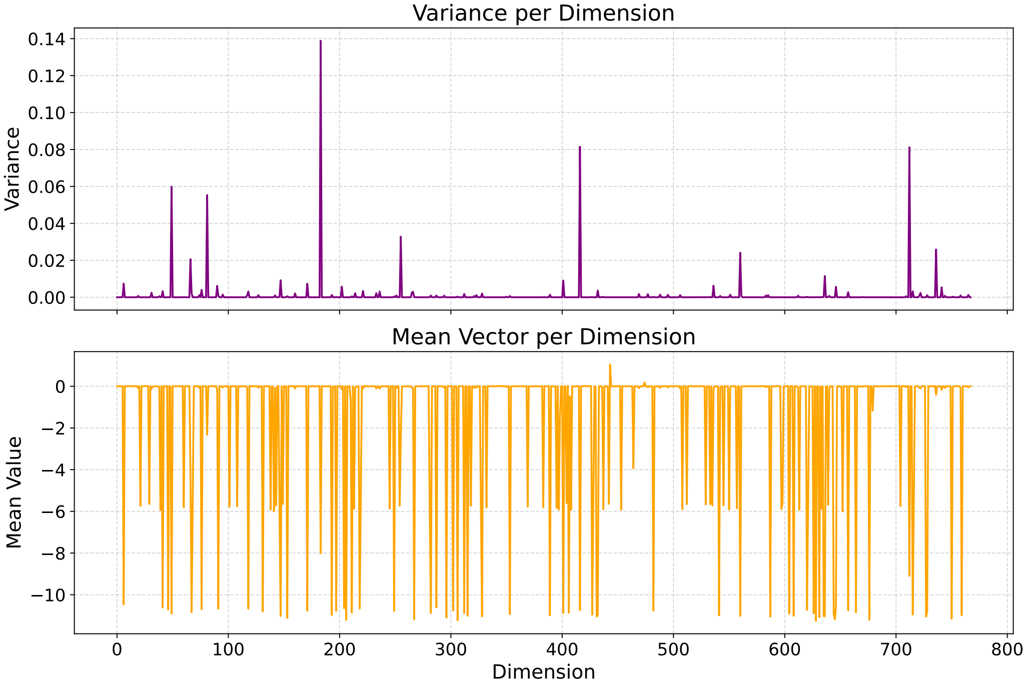


Figure 32: Variance per dimension of the conditional vector learned by MDSGen.

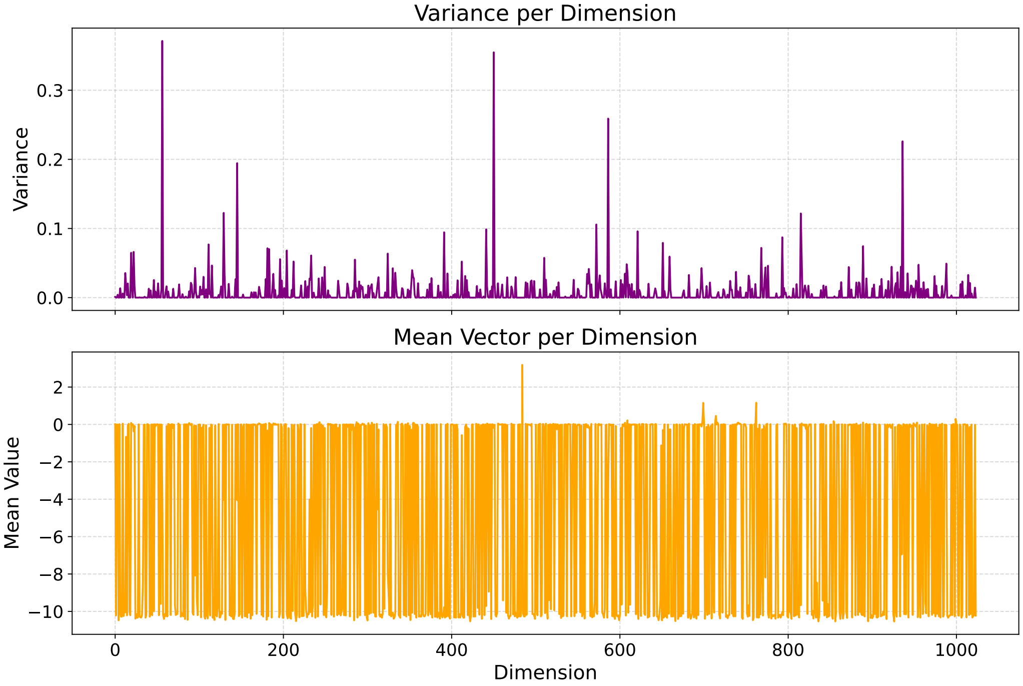


Figure 33: Variance per dimension of the conditional vector learned by X-MDPT.

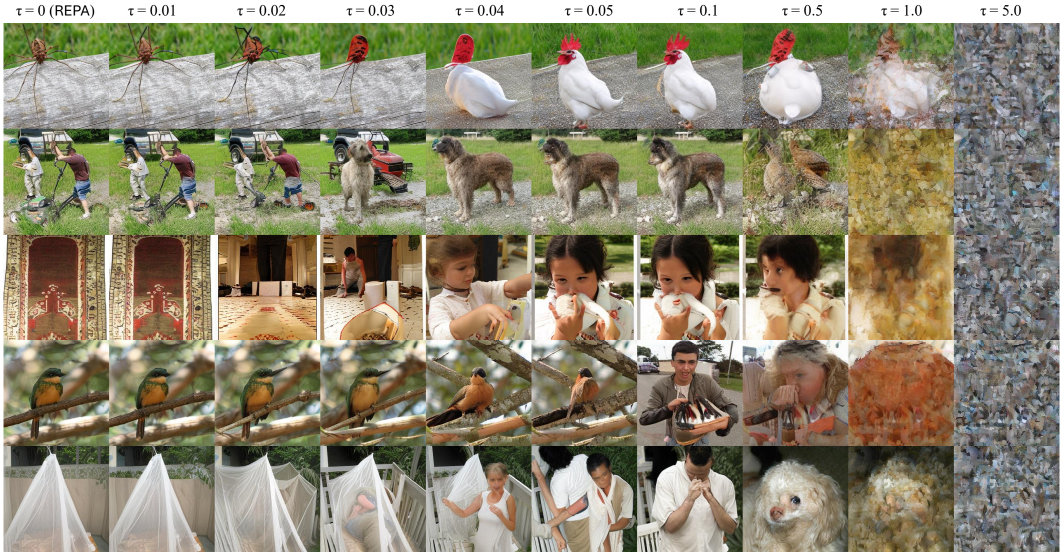


Figure 34: **Class-conditional ImageNet generation with pruned embeddings (1).** Removing low-magnitude dimensions from $\vec{c}$ preserves or slightly improves image quality, confirming that semantic information is concentrated in a few head dimensions.

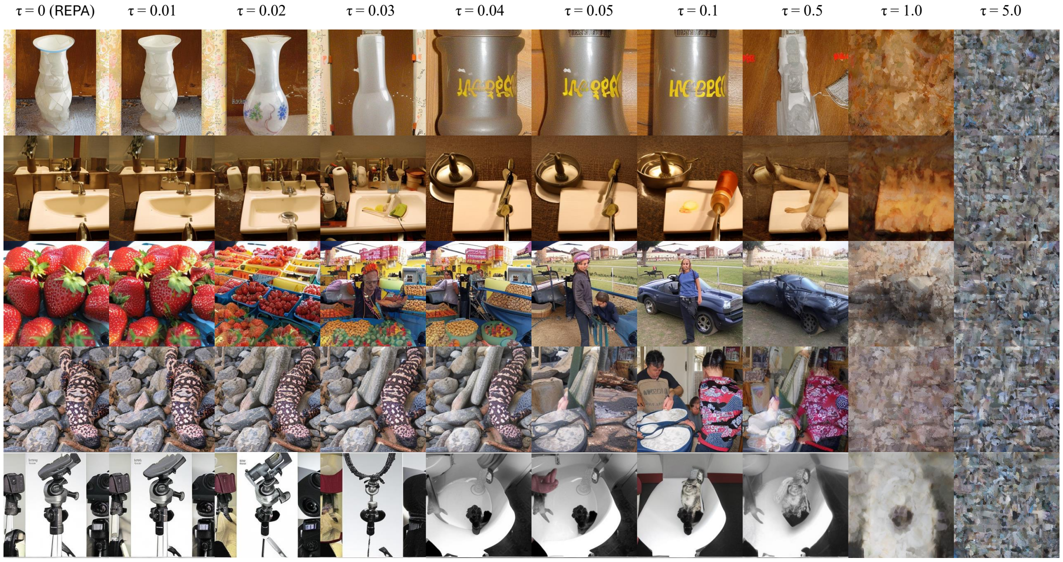


Figure 35: **Class-conditional ImageNet generation with pruned embeddings (2).** Removing low-magnitude dimensions from $\vec{c}$ preserves or slightly improves image quality, confirming that semantic information is concentrated in a few head dimensions.

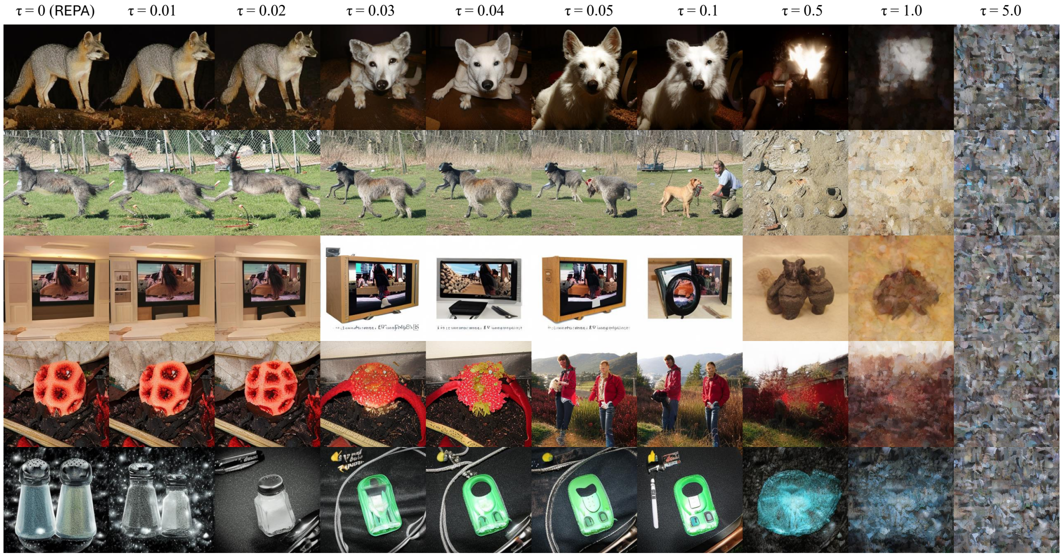


Figure 36: **Class-conditional ImageNet generation with pruned embeddings (3).** Removing low-magnitude dimensions from $\vec{c}$ preserves or slightly improves image quality, confirming that semantic information is concentrated in a few head dimensions.

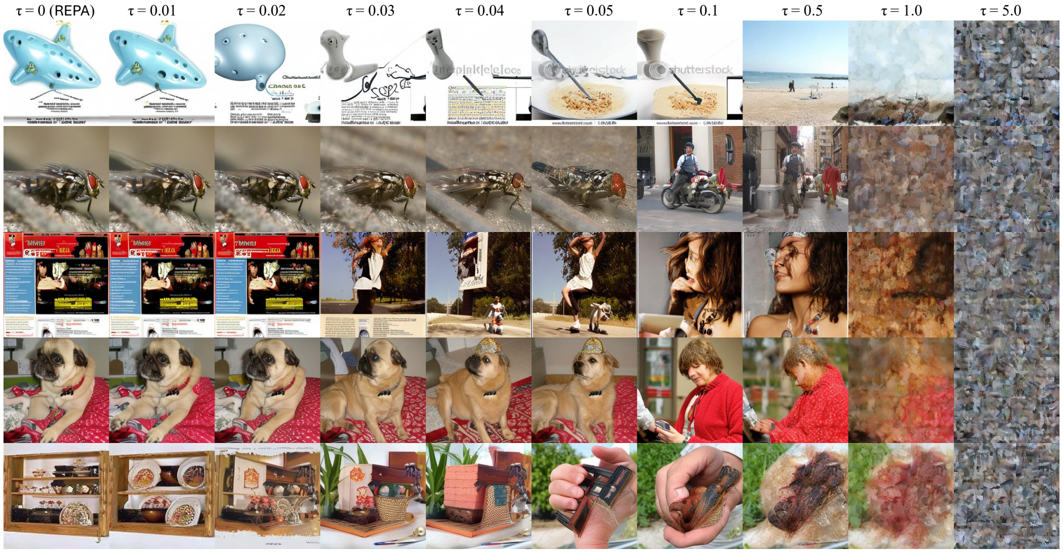


Figure 37: **Class-conditional ImageNet generation with pruned embeddings (4).** Removing low-magnitude dimensions from $\vec{c}$ preserves or slightly improves image quality, confirming that semantic information is concentrated in a few head dimensions.

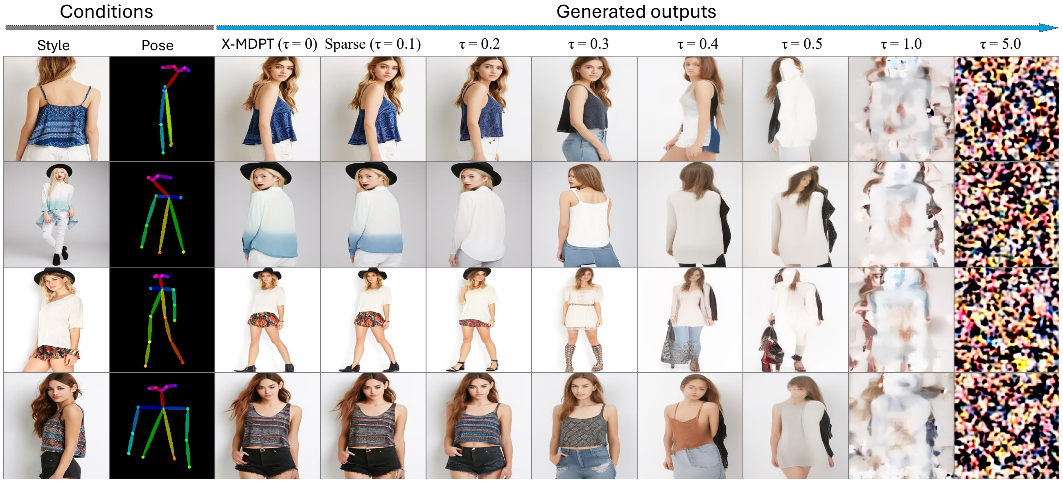


Figure 38: **Pose-guided person image synthesis with pruned embeddings (1).** Pruning tail dimensions in $\vec{c}$ maintains pose fidelity and visual quality, highlighting the redundancy of low-magnitude embedding dimensions.

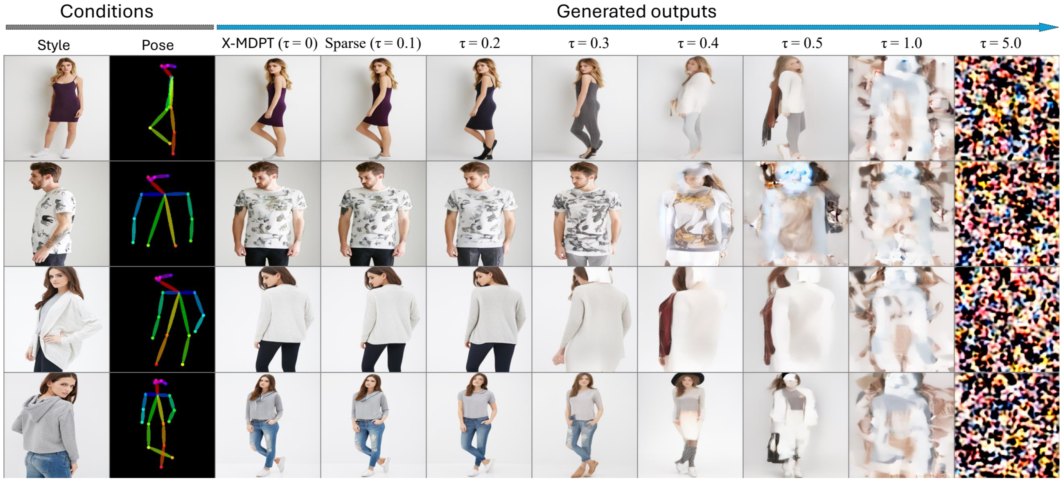


Figure 39: **Pose-guided person image synthesis with pruned embeddings (2).** Pruning tail dimensions in $\vec{c}$ maintains pose fidelity and visual quality, highlighting the redundancy of low-magnitude embedding dimensions.

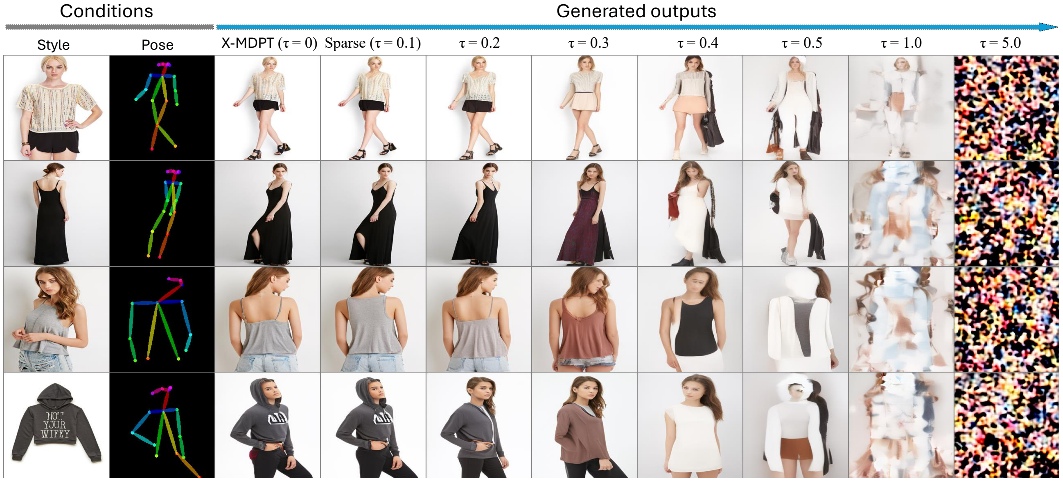


Figure 40: **Pose-guided person image synthesis with pruned embeddings (3).** Pruning tail dimensions in $\vec{c}$ maintains pose fidelity and visual quality, highlighting the redundancy of low-magnitude embedding dimensions.

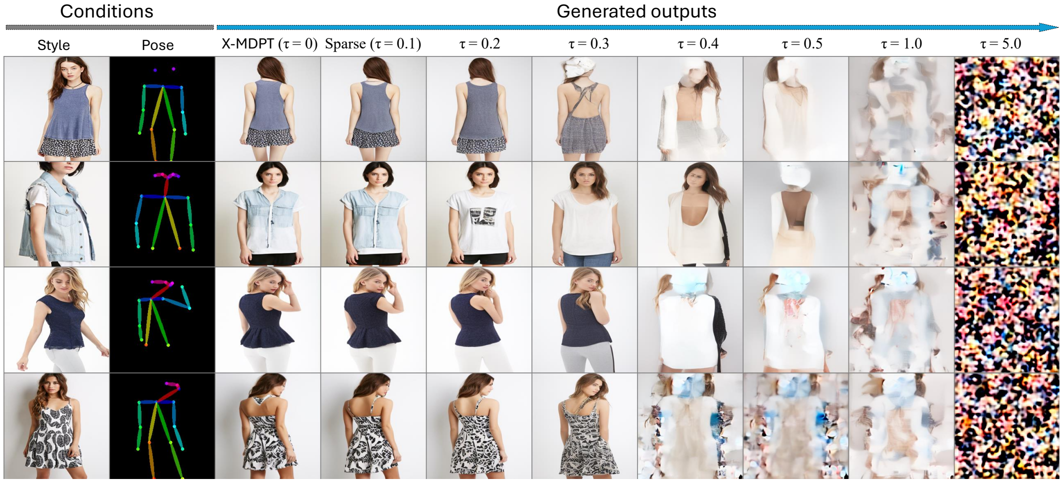


Figure 41: **Pose-guided person image synthesis with pruned embeddings (4).** Pruning tail dimensions in $\vec{c}$ maintains pose fidelity and visual quality, highlighting the redundancy of low-magnitude embedding dimensions.

B More Baselines and Analysis

B.1 Separating Timestep Embedding and Conditions Analysis.

Table 4: **Separate timestep (t) and conditions (y).** Participation Ratio (PR) in learned conditional embeddings of state-of-the-art models on Imagenet-1K class-conditioned generation. With ^{1} denotes the methods used: **AdaLN**, and ^{2} denotes the method used: **concatenation**.

Metrics	**DiT**^{1}	**SiT**^{1}	**MDT**^{1}	**LightningDiT**^{1}	**MG**^{1}	**REPA**^{1}	**UViT**^{2}	**Embed.**
Cosine Sim.	**0.9001**	**0.9852**	**0.9905**	**0.9779**	**0.9934**	**0.9946**	**0.97917**	y + t
nPR (α_{norm}) %	10.47	2.28	1.60	2.05	1.73	1.43	50.06	y + t
Cosine Sim.	0.7774	0.5436	0.8540	0.7166	0.6853	0.5194	0.00165	y
nPR (α_{norm}) %	70.14	37.43	36.75	36.42	43.67	41.60	63.52	y

B.2 Text-conditioned Methods and Model Sizes.

Table 5: **Separate timestep (t) and conditions (y).** Participation Ratio (PR) in learned conditional embeddings of state-of-the-art models on text or video-conditioned generation. With ^{1} denotes the methods used: **AdaLN**, and ^{3} denotes the method used: **cross-attention**.

Metrics	**X-MDPT-L**^{1}	**X-MDPT-B**^{1}	**X-MDPT-S**^{1}	**SD3.0 (2B)**^{1}	**SD3.0 (8B)**^{1}	**MDSGen**^{1}	**AudioLDM**^{3}	**Embed.**
Cosine Sim.	**0.9998**	**0.99992**	**0.9995**	**0.9962**	**0.9995**	**0.9999**	**0.9828**	y + t
nPR (α_{norm}) %	48.42	37.59	53.41	54.79	26.67	13.57	8.62	y + t
Cosine Sim.	**0.9862**	**0.9909**	**0.9492**	**0.9949**	**0.9937**	**0.9918**	0.1406	y
nPR (α_{norm}) %	20.68	29.75	37.51	52.39	24.25	9.98	63.09	y

B.3 More Quantitative Metrics.

Table 6: **Precision and Recall** with previous metrics: FID, IS, and CLIP.

Method	**FID**↓	**IS**↑	**CLIP**↑	**Precision**↑	**Recall**↑	**Remark**
REPA (Yu et al., 2025)	7.1694	**176.02**	29.746	0.8032	0.6236	Baseline
Pruned ($\tau = 0.01$) t_0	**7.1690**	175.97	**29.807**	0.7878	**0.6252**	**Ours**
Pruned ($\tau = 0.01$) $t_{n-k,n}$	**7.1598**	175.49	**29.805**	**0.8045**	**0.6381**	**Ours**
Model-Guide (Tang et al., 2025)	7.2478	174.5151	**30.199**	0.7842	0.6633	Baseline
Pruned ($\tau = 0.01$) t_0	**7.2466**	**174.5537**	**30.199**	**0.7854**	0.6625	**Ours**
Pruned ($\tau = 0.01$) $t_{n-k,n}$	**7.2455**	174.3103	30.198	**0.7898**	**0.6644**	**Ours**
LightningDiT (Yao et al., 2025)	7.0802	169.8574	30.720	0.7928	0.6248	Baseline
Pruned ($\tau = 0.01$) t_0	**7.0712**	169.9164	**30.729**	0.7906	**0.6256**	**Ours**
Pruned ($\tau = 0.01$) $t_{n-k,n}$	**7.0745**	**169.9236**	**30.729**	**0.7935**	**0.6265**	**Ours**

Table 7: Quantitative metrics on the DeepFashion dataset of pose-guide person image generation task with masked diffusion transformers.

Method	**FID** ↓	**SSIM**↑	**LPIPS**↓	**PSNR**↑	**Remark**
X-MDPT (Pham et al., 2024)	18.6372	0.6798	0.1672	17.336	Baseline
Pruned ($\tau = 0.1$) 40%	18.6692	0.6792	0.1675	17.328	**Ours**

B.4 Computational Reduction Analysis.

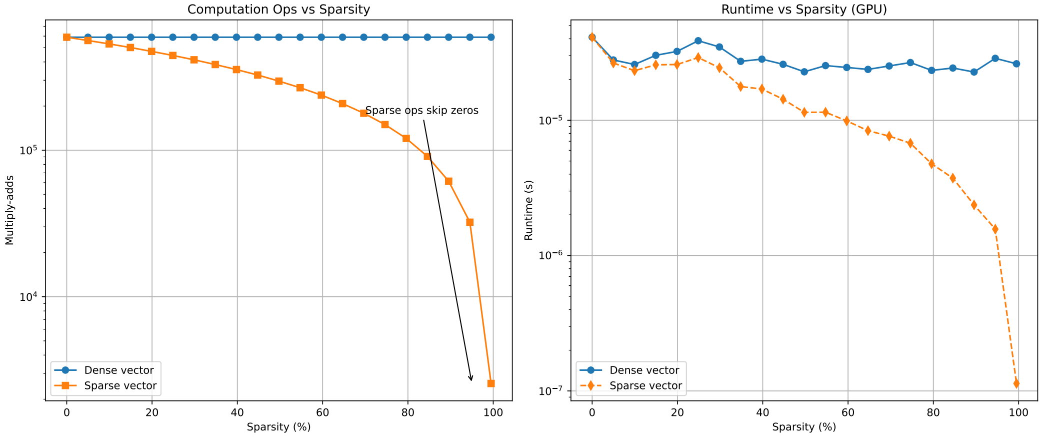


Figure 42: **Dense vs. Sparse vectors.** Compared the computation overhead. It shows that a sparse vector is more efficient in computation and has faster runtime than a dense vector (baseline).

B.5 Stability Analysis with Limited Timestep.

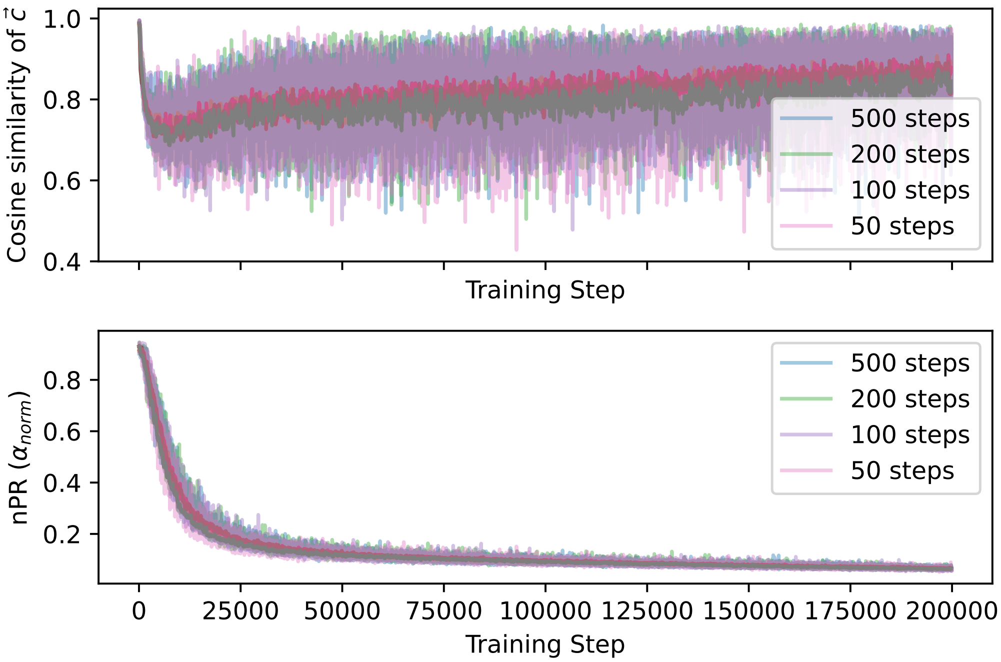


Figure 43: **Training dynamics of conditional embeddings.** Top: batchwise cosine similarity of $\vec{c}$ during training. Bottom: participation ratio (nPR) over training steps, showing progressive sparsification. **With limited timesteps: 50, 100, 200, and 500.**