A Theoretical Analysis of Curriculum Training in Diffusion Models

Anonymous Authors[1]

Abstract

Training a diffusion model (DM) amounts to solving a continuum of denoising subproblems whose difficulty varies with the noise level. The default practice samples these subproblems uniformly and updates all parameters jointly, which entangles feature representations and degrades generation. Recent empirical work suggests that presenting subproblems in an easy-to-hard order helps, yet why it helps and whether further gains are possible remain open. We propose a curriculum-based framework that schedules two ingredients during optimization: the difficulty of the denoising subproblems, and the share of parameters that is allowed to evolve. Training starts from easier subproblems on a subset of trainable neurons; harder subproblems are introduced later, with the remaining neurons gradually unlocked. The reserved capacity protects subtle, low-amplitude features from being overwritten while the network is still fitting coarse structure. We provide the first analysis connecting this curriculum to the training dynamics and generalization error of DMs, identifying a coarse-to-fine learning order that emerges implicitly under our schedule. Experiments on multiple datasets and architectures confirm the predicted gains over uniform-noise training.

1. Introduction

Diffusion Models (DMs) have emerged as a dominant paradigm for generative modeling, achieving state-of-the-art performance across various applications, including image synthesis (Ho et al., 2020; Song et al., 2020; Rombach et al., 2022), video generation (Ho et al., 2022), and multimodal content creation (Ramesh et al., 2022). The training of diffusion models consists of a forward noising process that progressively adds noise to clean data and a reverse process that trains a neural network to predict and remove noise. As a result, diffusion training requires learning denoising functions across a continuum of noise levels, ranging from heavily corrupted inputs to nearly clean data. Typically, training samples are drawn uniformly across all noise levels, leading the model to learn all denoising tasks simultaneously. This can be challenging because high-noise and low-noise denoising tasks differ substantially in their difficulty and targeted features: high-noise denoising recovers coarse global structure, while low-noise denoising refines fine-grained details.

To address the ineffectiveness of uniform sampling, a line of work focuses on structuring the denoising objectives across various noise levels. These methods operate either by reweighting the training loss to emphasize specific noise regimes (Hang et al., 2023) or explicitly scheduling the noise levels encountered during training (Hoogeboom et al., 2023). A key empirical observation is that denoising at high noise levels primarily captures global structure and semantic information, whereas low-noise denoising focuses on fine-grained details and textures (Karras et al., 2022; Choi et al., 2022). Built upon this finding, a recent study (Kim et al., 2024) introduces a denoising schedule that progresses from high-noise to low-noise, empirically observing faster convergence and improved sample quality compared to standard uniform noise sampling.

Despite strong empirical performance, why high-to-low denoising schedules improve diffusion training remains theoretically unexplored; even standard diffusion training with uniform noise sampling is poorly understood. Existing analyses typically separate optimization from generalization, either characterizing optimal solutions without showing how they are obtained (Block et al., 2020; Oko et al., 2023) or analyzing convergence without connecting to learning performance (Lee et al., 2023; Chen et al., 2023; Conforti et al., 2023). Recent work begins to examine internal training dynamics, including representation dynamics (Li et al., 2025), coarse-to-fine sampling (Wang & Vastola, 2023; Wang & Pehlevan, 2025), and feature learning (Han et al., 2024). However, these results are restricted to over-simplified linear models and uniform noise sampling. Several key questions remain to be addressed:

- **Theoretical training dynamics and generalization:** How does noise level scheduling shape learning perfor-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

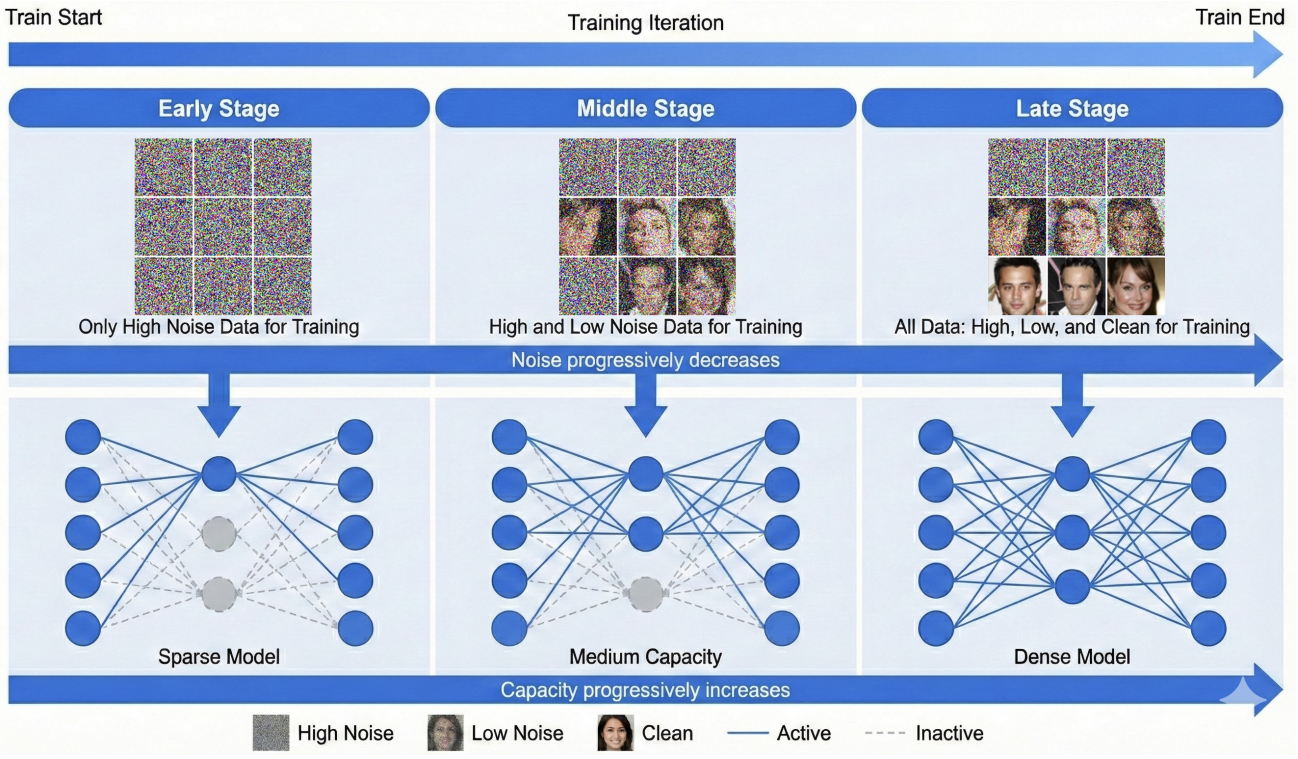

Figure 1. **Illustration of joint curriculum high-to-low denoising schedule and model sparsity for diffusion training.** Our framework coordinates two high-to-low progressions: noise levels decrease from high to low, and model sparsity simultaneously decreases by activating more neurons. In the early stage, only high-noise data is used with a sparse model; in the late stage, all noise levels, including clean data, are used with a fully activated model.

mance? Can we provide formal guarantees for the benefits of high-to-low denoising schedules over uniform training?

- **Beyond denoising schedule:** Can performance be further improved through strategies beyond noise scheduling?

This paper addresses these questions by developing a joint high-to-low denoising schedule and model sparsity framework for diffusion training, and providing the first theoretical analysis of how these strategies improve feature learning and generalization. As illustrated in Fig. 1, our framework coordinates high-to-low progressions in two orthogonal dimensions: the noise level decreases from high to low while model sparsity decreases from sparse to dense via progressive neuron activation. We analyze the resulting training dynamics on a two-layer ReLU network, which is more general than the linear models in most prior diffusion theory. Our main contributions are as follows.

1. **Theoretical characterization of standard diffusion training with uniform sampling and its limitations**. Our analysis shows that diffusion training learns features in a stage-wise manner: coarse features are acquired in early stages, while fine-grained features emerge in later stages. In standard training, high- and low-noise samples are mixed throughout optimization, causing weak fine-grained features to become entangled with other features and high-noise inputs. This prevents the formation of purified representations and results in lower-quality generation of subtle features.

2. **Theoretical justification of high-to-low denoising schedule**. We demonstrate that high-to-low denoising schedule naturally aligns with the intrinsic feature learning order of diffusion models. By presenting high-noise samples early and low-noise samples later, coarse features, which are robust to high noise, are learned first, while fine-grained features are acquired under low-noise conditions in later stages. This stage-wise alignment significantly improves fine-feature quality and enhances generalization. As shown in Fig. 2 (Denoise-Sched. only), the denoising schedule achieves lower FID at the same training FLOPs and produces samples with better-preserved fine details compared to standard training.

3. **Joint denoising schedule and model sparsity framework with theoretical and empirical support**. Beyond the high-to-low denoising schedule, our proposed joint method additionally introduces high-to-low model sparsity that gradually activates neurons during training. This reserves part of the network for learning fine-grained features in later stages, preventing interference from early high-noise training and enabling purified representations of fine-grained features. Fig. 2 shows that our joint method achieves the lowest FID with the fewest training FLOPs and the best image generation quality at the same final training step on the CelebA-64

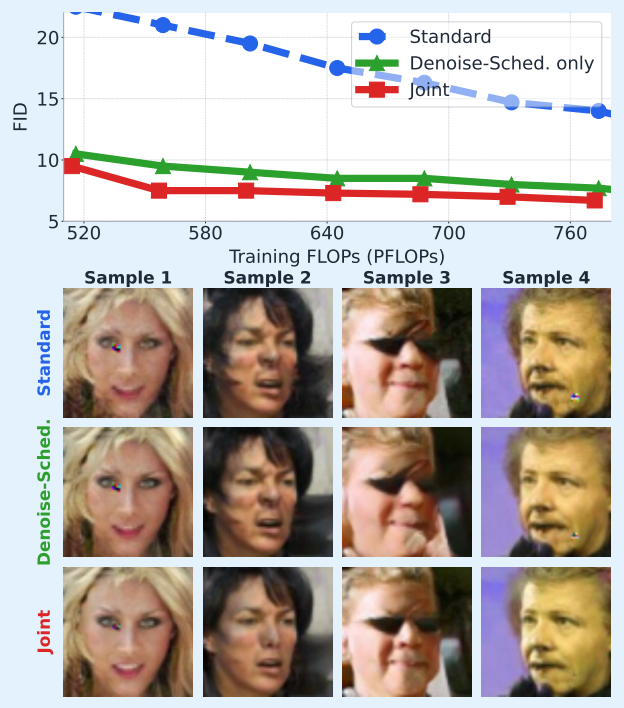

Figure 2. FID vs. training FLOPs on CelebA-64 using U-ViT (Bao et al., 2023). **Top**: FID curves against cumulative training FLOPs. **Bottom**: generated samples at the same final training step (200k) for all three methods. **Standard**: uniform noise sampling. **Denoise-Sched. only**: high-to-low denoising schedule only. **Joint**: high-to-low denoising schedule with high-to-low model sparsity. Our joint method achieves the lowest FID with the fewest training FLOPs, and yields fewer distorted details and sharper fine-grained features at the same final training step. More samples in Appendix A.2 and A.4.

dataset.

1.1. Related Works

Theoretical Analyses of Neural Networks. The neural tangent kernel approach (Jacot et al., 2018; Allen-Zhu & Li, 2022) shows that overparameterized networks are well approximated by a linear model around initialization, under which gradient-based training converges to a near-optimal solution. The feature learning framework (Li et al., 2025; Wen & Li, 2021) instead assumes data is a combination of underlying features and analyzes how shallow networks progressively identify them during training, yielding strong generalization.

Diffusion Model: Training and Architecture. The denoising network is typically a UNet (Ronneberger et al., 2015), with recent works (Peebles & Xie, 2023; Bao et al., 2023) exploring scalable Transformer architectures. Latent Diffusion Models (Rombach et al., 2022) run diffusion in a compressed latent space to cut cost. Beyond denoising schedules, other training strategies include synthetic-to-real complexity progression (Liang et al., 2025) and preference pair difficulty scheduling (Croitoru et al., 2025).

2. Diffusion Model

Diffusion models learn to reverse a data corruption process by training a neural network to denoise noisy observations across multiple noise levels. During training, clean data $x_0 \in \mathbb{R}^{d_1}$ is corrupted by additive Gaussian noise,

$$x_\tau = x_0 + \tau\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I_{d_1}), \tag{1}$$

where $\tau > 0$ is the noise level. This corruption induces a noisy marginal distribution $p_\tau(x_\tau) = \int p_0(x_0)\mathcal{N}(x_\tau \mid x_0, \tau^2 I)\, dx_0$, whose score function $\nabla_{x_\tau} \log p_\tau(x_\tau)$ is the target of learning.

Following (Vincent, 2011; Song & Ermon, 2019), we aim to learn a function $s(x_\tau, \tau)$ to approximate the score $\nabla_{x_\tau} \log p_\tau(x_\tau)$. Since the true score is intractable, denoising score matching replaces it with the tractable conditional score $\nabla_{x_\tau} \log p_\tau(x_\tau|x_0) = -(x_\tau - x_0)/\tau^2$. Under Gaussian noise, the score admits a closed expression, which allows us to reparameterize it via a denoiser g_θ as

$$s(x_\tau, \tau) = (g_\theta(x_\tau) - x_\tau)/\tau^2. \tag{2}$$

This further leads to the population denoising objective as

$$\mathbb{E}_{x_0 \sim p_{\text{data}}, \epsilon} \|g_\theta(x_\tau) - x_0\|_2^2. \tag{3}$$

Given a training set $\mathcal{D}$ of i.i.d. clean samples $\{x_{0,j}\}_{j=1}^N$ where $x_{0,j} \sim p_{\text{data}}$. For each fixed clean sample $x_{0,j}$, we have a collection of noisy data generated by the forward process, denoted as $x_{\tau,j}$. For notational convenience, when referring to a generic data sample, we omit the sample index j and write $x_{\tau,j}$ simply as x_τ. Then, diffusion training optimizes over θ to minimize the empirical risk

$$\widehat{L}_{DM}(g) := \frac{1}{|\mathcal{D}|} \sum_{x_0 \in \mathcal{D}} \mathbb{E}_{\tau,\epsilon} \frac{1}{2} \|g_\theta(x_\tau) - x_0\|_2^2. \tag{4}$$

After training, sample generation starts from random noise and proceeds through iteratively applying Langevin dynamics guided by the learned score model. After a sufficient number of iterations, the generated samples approximate clean data drawn from p_{data}.

3. Joint Denoising-Sparsity Scheduling

A diffusion model is trained via stochastic gradient descent (SGD) with step size η and batch size B to minimize the empirical risk in (4). Under **standard training**, the noise level τ is sampled uniformly from a fixed interval $[\tau_{\min}, \tau_{\max}]$ throughout optimization, and all model parameters are updated at every iteration (i.e., no sparsity). This exposes the network to the full noise range without any temporal structure.

Recent **high-to-low denoising schedule only** approaches instead schedule the noise level from high to low during training, while keeping all parameters active (i.e., no sparsity),

so effective task difficulty grows over time. Higher noise levels are considered easier, since early training mainly recovers coarse, large-scale structure rather than fine details. As lower noise levels are introduced, the reconstruction must become increasingly fine-grained, making later tasks harder. Such scheduling improves generalization and convergence empirically (Kim et al., 2024), but its theoretical foundations remain underexplored.

This paper provides a theoretical understanding of noise scheduling in training dynamics. Guided by our analysis, we propose a **joint high-to-low denoising schedule and model sparsity framework** that jointly schedules the noise level and effective model sparsity. Training is organized into multiple stages with progressively decreasing noise level and increasing active parameters. Early stages focus on coarse, high-noise tasks using a restricted parameter subset, while later stages introduce low-noise reconstruction requiring finer accuracy and more parameters. Specifically, for M-stage training, the noise range $[\tau_{\min}, \tau_{\max}]$ is divided into M intervals $[\tau_{i-1}, \tau_i]$ for $i \in [M]$, with $\tau_{\min} = \tau_0 < \tau_1 < \cdots < \tau_M = \tau_{\max}$, and parameters expand as $\theta_1 \subseteq \theta_2 \subseteq \cdots \subseteq \theta_M = \theta$. In stage i, τ is sampled from $[\tau_{i-1}, \tau_i]$ and only θ_i is updated. If θ is fully updated across all M stages, our method reduces to existing noise-only high-to-low schedules.

4. Main Theoretical Results

4.1. Key Theoretical Insights

Before presenting the formal theoretical setup and results, we summarize the main insights of our analysis. We consider a signal model in which clean data contain both coarse features, represented by dictionary $\mathbf{M}_1$ and magnitude α_1, and fine-grained features, represented by dictionary $\mathbf{M}_2$ and magnitude α_2, with $\alpha_1 \gg \alpha_2$. The maximum noise level satisfies $\alpha_2 < \tau_{\max} < \alpha_1$. We study a two-stage joint denoising-sparsity scheduling that partitions the noise range $[\tau_{\min}, \tau_{\max}]$ at a threshold $\tau_1 < \alpha_2$. Our analysis reveals three key findings.

(I) Joint denoising-sparsity scheduling yields pure, disentangled representations. In Stage I (higher noise), since $\tau_{\max} < \alpha_1$, the network reliably learns coarse features. For each feature direction in $\mathbf{M}_1$, a constant number of neurons specialize in that direction and achieve near-perfect alignment. In Stage II (lower noise), since $\tau_1 < \alpha_2$, and $\mathbf{M}_1$ is already accurately learned, a separate subset of neurons can align with individual fine-grained features in $\mathbf{M}_2$ without interference. Consequently, every feature in $\mathbf{M}_1$ and $\mathbf{M}_2$ has pure neuron-wise representations.

For comparison, high-to-low denoising schedule only can align with features in $\mathbf{M}_1$ and $\mathbf{M}_2$, but because no neurons are frozen in Stage I, noise-induced interference accumulates and degrades the purity of representations for $\mathbf{M}_2$.

(II) Standard training produces mixed, entangled representations. Without high-to-low scheduling, noise is sampled up to $\tau_{\max}$ throughout training. While coarse features in $\mathbf{M}_1$ can be learned accurately, fine-grained features in $\mathbf{M}_2$ cannot be isolated. Neurons encoding $\mathbf{M}_2$ inevitably mix multiple feature directions and noise components, yielding entangled representations and imperfect recovery of fine-grained structure.

(III) High-to-low benefits appear in low-noise generalization. Both high-to-low methods and standard training achieve comparable generalization performance at high noise levels. However, in the low-noise regime, where reconstruction error is dominated by fine-grained features $\mathbf{M}_2$, joint denoising-sparsity scheduling maintains small generalization error, whereas standard training incurs order-wise larger errors. This explains the empirical observation that high-to-low scheduling particularly benefits sample quality at final generation steps.

4.2. Definitions and Assumptions

Denoising Network Architecture: We use a one-hidden-layer ReLU network as the denoiser, which serves as a tractable model for analyzing training dynamics while capturing the nonlinearity present in practical architectures. The notation summary is given in Appendix B.

Definition 1 (Denoising network)**.** The denoiser $g : \mathbb{R}^{d_1} \to \mathbb{R}^{d_1}$ is defined as

$$g(x) = V^\top \sigma(Wx - b), \tag{5}$$

where $W = [w_1, \ldots, w_m]^\top \in \mathbb{R}^{m \times d_1}$ and $V = [v_1, \ldots, v_m]^\top \in \mathbb{R}^{m \times d_1}$ are the weight matrices, $b = (b_1, \ldots, b_m)^\top \in \mathbb{R}^m$ is the bias vector with b_i acting as the activation threshold for the i-th hidden neuron, and σ denotes the ReLU activation function.

Existing theoretical analyses of diffusion models are largely limited to linear networks (Wang & Pehlevan, 2025), infinite-width networks (Boffi et al., 2024), or assume an optimal denoiser without analyzing training dynamics (Li et al., 2024). More broadly, theoretical studies of feature learning in neural networks have focused on one-hidden-layer networks in supervised and self-supervised settings (Allen-Zhu & Li, 2022; Wen & Li, 2021). We extend the analysis to nonlinear ReLU networks trained via SGD, characterizing how parameters evolve during training.

Training Setup: We analyze a two-stage joint denoising-sparsity schedule for clarity; the analysis extends to M stages. The weight matrices W and V are initialized with Gaussian entries $w_i^{(0)}, v_i^{(0)} \sim \mathcal{N}(0, \sigma_0^2 I_{d_1})$ and updated by SGD with step size η and batch size B. Biases are initialized at $b_i^{(0)} = \Theta(\sigma_0 \sqrt{\log d})$ and grown jointly during training, following the bias-growth schedule of (Wen & Li, 2021). In

Stage I ($t \in [0, T_1]$), $\tau \sim \mathrm{Unif}[\tau_1, \tau_{\max}]$, and only weight pairs (w_i, v_i) with i in a randomly chosen subset $S_1 \subset [m]$ are updated; the remaining neurons stay frozen at initialization. In Stage II ($t \in [T_1, T_1 + T_2]$), $\tau \sim \mathrm{Unif}[\tau_{\min}, \tau_1]$ and all weights are updated. For comparison, we also analyze standard training, where $\tau \sim \mathrm{Unif}[\tau_{\min}, \tau_{\max}]$ and all weights are updated throughout.

Data Model: We adopt the multi-scale sparse coding model in Def. 2 for the clean data. We consider two feature-strength scales for simplicity; the model naturally extends to multiple scales. Specifically,

Definition 2 (Multi-scale sparse coding model)**.** The clean signal $x_0 \in \mathbb{R}^{d_1}$ admits a sparse decomposition:

$$x_0 = \alpha_1 \mathbf{M}_1 z_1 + \alpha_2 \mathbf{M}_2 z_2, \tag{6}$$

where $x_0 \in \mathbb{R}^{d_1}$, $z_1, z_2 \in \mathbb{R}^d$, and $d_1 = \mathrm{poly}(d)$, we have:

(a) **Orthonormal dictionaries:** $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{d_1 \times d}$ are column-orthonormal matrices, representing two disjoint feature subspaces. We write $\mathbf{M}_k = [\mathbf{M}_{k1}, \ldots, \mathbf{M}_{kd}]$ for $k \in \{1, 2\}$, where $\mathbf{M}_{kj} \in \mathbb{R}^{d_1}$ denotes the j-th column of $\mathbf{M}_k$.

(b) **Sparse latent codes:** $z_1, z_2 \in \mathbb{R}^d$ are sparse vectors with $z_{k,j} \in \{0, \pm 1\}$ and $|z_{k,j}| \sim \mathrm{Bernoulli}(C_z/d)$ for some constant $C_z > 0$.

(c) **Scale separation:** $\alpha_1 = \Theta(1) \gg \alpha_2 = \Theta(1/d^{c_0})$ for constant $c_0 \in (0, 1)$.

The scale separation $\alpha_1 \gg \alpha_2$ models the natural hierarchy in visual data: $\mathbf{M}_1$ represents dominant, coarse-grained features (e.g., object shapes, global structure) with amplitude $\alpha_1 = \Theta(1)$, while $\mathbf{M}_2$ captures weaker, fine-grained details (e.g., textures, local patterns) with amplitude $\alpha_2 = \Theta(1/d^{c_0})$.

This model generalizes standard sparse coding, which uses a single-scale dictionary. Sparse coding underlies many theoretical studies, including supervised (Allen-Zhu & Li, 2022) and self-supervised learning (Wen & Li, 2021; Pareek et al., 2025), and diffusion models (Boffi et al., 2024; Wang & Pehlevan, 2025; Li et al., 2025). The multi-scale design reflects the spectral structure of images: (Pope et al., 2021; Kadkhodaie et al., 2024) show images decompose into a sparse orthonormal basis whose dominant eigenvalues capture low-frequency coarse structure (analogous to $\mathbf{M}_1$) and smaller eigenvalues capture high-frequency fine details (analogous to $\mathbf{M}_2$).

4.3. Feature Alignment with Different Training Methods

We analyze the training dynamics and generalization of three training methods: (i) joint denoising-sparsity scheduling, (ii) high-to-low denoising schedule only, and (iii) standard training. These methods can all learn $\mathbf{M}_1$ accurately (Theorem 1) but differ in learning $\mathbf{M}_2$ (Theorems 2-4), resulting in different generative quality (Theorem 5). The proof sketch is provided in Appendix B.2.

Condition 1. $m = d^{1.1}$, $T_1 = \Theta\Big(\frac{d \log d}{\eta}\Big)$, $T_2 = \Theta\Big(\frac{d^{1+2c_0} \log d}{\eta}\Big)$, $\tau_{\min} = \Theta(1/d_1)$, $\tau_{\max} = \Theta(\alpha_1/\sqrt{\log d})$, $\tau_1 = \Theta(\alpha_2/\sqrt{\log d})$, $|S_1| = O(d^{1.01})$.

Under these parameter settings in Condition 1, $\tau_{\max} < \alpha_1$, so the $\mathbf{M}_1$ features remain stronger than the highest noise and can always be accurately recovered. On the other hand, $\tau_{\max} > \alpha_2$, which means the $\mathbf{M}_2$ features are weaker than the highest noise, making their recovery inaccurate if learned using standard training that samples noise uniformly across all levels. By setting $\tau_1 = \Theta(\alpha_2/\sqrt{\log d}) < \alpha_2$, in Stage II of high-to-low training the noise is restricted to $[\tau_{\min}, \tau_1]$, which is smaller than the $\mathbf{M}_2$ signal, allowing accurate recovery of $\mathbf{M}_2$. We next formalize this initialization as follows.

Theorem 1 (**Accurate $\mathbf{M}_1$-Alignment by Three Training Protocols**)**.** *Assume Condition 1 holds. For all three training methods, for every $j \in [d]$, there exist at least $\Theta(1)$ neurons that achieve pure $\mathbf{M}_{1j}$-alignment, i.e., for each such neuron i, for all $t > T_1$,*

$$w_i^{(t)} = \alpha_{i,j} \mathbf{M}_{1j} + \mathbf{r}_i, \quad v_i^{(t)} = \alpha_{i,j} \mathbf{M}_{1j} + \mathbf{s}_i, \quad \forall t > T_1 \tag{7}$$

where $\alpha_{i,j}^2 = (1 - o(1))(\|w_i^{(t)}\|_2^2 + \|v_i^{(t)}\|_2^2)/2$. Here $\mathbf{r}_i, \mathbf{s}_i \perp \mathrm{span}(\mathbf{M}_1)$ denote the residual errors and $\|\mathbf{r}_i\|_2^2, \|\mathbf{s}_i\|_2^2 = o(\|w_i^{(t)}\|_2^2 + \|v_i^{(t)}\|_2^2)$.

Remark 1. Theorem 1 shows all three training methods can learn individual features in $\mathbf{M}_1$ accurately. Because $\mathbf{M}_1$ model coarse features, even if high-to-low methods schedule higher noise at the early stages, $\mathbf{M}_1$ can still be learned accurately as long as the noise is not too high.

Theorem 2 (**Pure $\mathbf{M}_2$-Alignment with Joint Denoising-Sparsity Scheduling**)**.** *Assume Condition 1 holds. For every $j \in [d]$, there exist at least $\Theta(1)$ neurons that achieve pure $\mathbf{M}_{2j}$-alignment, i.e., for every such neuron i, for all $t > T_1 + T_2$,*

$$\hat{w}_i^{(t)} = \hat{\beta}_{i,j} \mathbf{M}_{2j} + \hat{\mathbf{r}}_i, \quad \hat{v}_i^{(t)} = \hat{\beta}_{i,j} \mathbf{M}_{2j} + \hat{\mathbf{s}}_i, \tag{8}$$

where $\hat{\beta}_{i,j}^2 = (1 - o(1)) \frac{\|\hat{w}_i^{(t)}\|_2^2 + \|\hat{v}_i^{(t)}\|_2^2}{2}$, and $\|\hat{\mathbf{r}}_i\|_2^2, \|\hat{\mathbf{s}}_i\|_2^2 = o(\|\hat{w}_i\|_2^2 + \|\hat{v}_i\|_2^2)$.

Remark 2. Theorem 2 shows joint scheduling accurately learns individual $\mathbf{M}_2$ features. Neurons frozen in Stage I ($t < T_1$) stay at initialization scale, accumulating no interference from dominant $\mathbf{M}_1$ features. Once unfrozen in Stage II, they specialize in individual $\mathbf{M}_2$ features with negligible residual error, yielding purified $\mathbf{M}_2$ representations.

Theorem 3 (**Impure $\mathbf{M}_2$-Alignment with High-to-Low Denoising Schedule Only**)**.** *Assume Condition 1 holds. For*

every $j \in [d]$ such that, there exist at least $\Theta(1)$ neurons that are partially aligned with $\mathbf{M}_{2j}$, i.e., for each such neuron i

$$\tilde{w}_i^{(t)} = \tilde{\beta}_{i,j}\mathbf{M}_{2j} + \tilde{\mathbf{r}}_i, \quad \tilde{v}_i^{(t)} = \tilde{\beta}_{i,j}\mathbf{M}_{2j} + \tilde{\mathbf{s}}_i, \quad \forall t > T_1 + T_2 \tag{9}$$

where $\tilde{\beta}_{i,j}^2 = \Theta(\|\tilde{w}_i^{(t)}\|_2^2 + \|\tilde{v}_i^{(t)}\|_2^2)$. Here $\tilde{\mathbf{r}}_i, \tilde{\mathbf{s}}_i \perp \mathrm{span}(\mathbf{M}_2)$ denote residual error with $\|\tilde{\mathbf{r}}_i\|_2^2, \|\tilde{\mathbf{s}}_i\|_2^2 = \Theta(\|\tilde{w}_i^{(t)}\|_2^2 + \|\tilde{v}_i^{(t)}\|_2^2)$.

Remark 3. Theorem 3 shows that high-to-low denoising scheduling alone can learn individual features in $\mathbf{M}_2$, but with non-negligible residual errors, degrading performance relative to joint scheduling. This is because neurons learning $\mathbf{M}_{2j}$ in Stage II accumulate non-negligible errors during Stage I updates, and these errors persist through training.

Theorem 4 (Entangled $\mathbf{M}_2$ learning under Standard Training). *Assume Condition 1 holds. For every $j \in [d]$ such that, there exist at least $\Theta(1)$ neurons i that learn $\mathbf{M}_{2j}$ mixed with other features in $\mathbf{M}_2$, i.e., for every such neuron i there exists $\mathcal{N}_i \subseteq [d]$ with $j \in \mathcal{N}_i$, and $|\mathcal{N}_i| \geq \Omega\left(\sqrt{d/\log d}\right)$ such that for $t \geq T_1$*

$$\bar{w}_i^{(t)} = \sum_{j' \in \mathcal{N}_i} \bar{\beta}_{i,j'}\mathbf{M}_{2j'} + \bar{\mathbf{r}}_i, \quad \bar{v}_i^{(t)} = \sum_{j' \in \mathcal{N}_i} \bar{\beta}_{i,j'}\mathbf{M}_{2j'} + \bar{\mathbf{s}}_i, \tag{10}$$

where $\bar{\beta}_{i,j'}^2 = \Theta((\|\bar{w}_i\|_2^2 + \|\bar{v}_i\|_2^2)/|\mathcal{N}_i|)$, residuals $\bar{\mathbf{r}}_i, \bar{\mathbf{s}}_i \perp \mathrm{span}(\mathbf{M}_2)$ satisfy $\|\bar{\mathbf{r}}_i\|_2^2, \|\bar{\mathbf{s}}_i\|_2^2 = \Theta(\|\bar{w}_i\|_2^2 + \|\bar{v}_i\|_2^2)$, and $|\langle \bar{w}_{i'}, \mathbf{M}_{2j}\rangle| = O((\|\bar{w}_i\|_2^2 + \|\bar{v}_i\|_2^2)/d)$ for all other neurons i'.

Remark 4. Theorem 4 shows that standard training fails to learn $\mathbf{M}_2$ features properly. Unlike Theorem 3 where neurons align to a single $\mathbf{M}_{2j}$ with a small residual, here neurons that learn $\mathbf{M}_{2j}$ would always learn some other features $\mathbf{M}_{2j'}$ as well, as shown in (10), leading to an entangled representation. That is because the highest noise level $\tau_{\max}$ in standard training is greater than $\mathbf{M}_2$ feature strength α_2, hurting the learning of $\mathbf{M}_2$ features.

Theorem 5 (Generalization across noise regimes)**.** *Let g_{joint} and g_{std} denote networks trained with joint scheduling and standard training, and define $\mathcal{L}(g;\tau) := \mathbb{E}_{x_0 \sim p_{\text{data}}, \tau, \epsilon}\|g(x_\tau) - x_0\|_2^2$.*

(a) For high noise $\tau \in [\tau_1, \tau_{\max}]$, both methods achieve the same generalization accuracy:

$$\mathcal{L}(g;\tau) = \Theta(1/d^{2c_0}). \tag{11}$$

(b) For low noise $\tau \in [\tau_{\min}, \tau_1]$, joint scheduling outperforms standard training:

$$\mathcal{L}(g_{\text{joint}};\tau) = O(1/d^2), \quad \mathcal{L}(g_{\text{std}};\tau) = \Theta(1/d^{2c_0}). \tag{12}$$

Remark 5. In the high-noise regime, $\mathbf{M}_1$ features are recovered accurately while $\mathbf{M}_2$ features cannot be recovered, as noise dominates the weak $\mathbf{M}_2$. Since all three protocols successfully learn $\mathbf{M}_1$ (Theorem 1), they perform identically here. In the low-noise regime, the three methods differ in their learned $\mathbf{M}_2$ representations: joint scheduling learns pure $\mathbf{M}_2$ features (Theorem 2) and achieves full recovery; denoising scheduling alone learns impure $\mathbf{M}_2$ features with residual contamination (Theorem 3) and achieves partial recovery; standard training fails to learn $\mathbf{M}_2$ (Theorem 4) and cannot recover them.

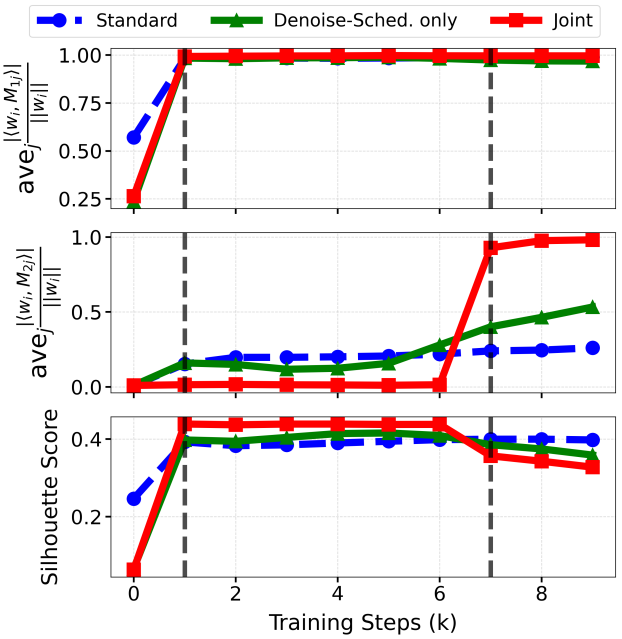

Figure 3. Feature alignment and Silhouette Score dynamics. **Top**: $\mathbf{M}_1$ feature ratio. **Middle**: $\mathbf{M}_2$ feature ratio. **Bottom**: Silhouette Score. Dashed lines mark phase transitions. SS dynamics closely track feature alignment, confirming that SS reflects feature purity.

5. Experiments

5.1. Synthetic Experiments

Experiment Setup. We validate our theoretical predictions on the multi-scale sparse coding model. Data is generated as $x_0 = \alpha_1\mathbf{M}_1 z_1 + \alpha_2\mathbf{M}_2 z_2$ with orthonormal dictionaries $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{d_1 \times d}$ and z_1, z_2 having i.i.d. Bernoulli(0.1) entries. Network weights $W, V \in \mathbb{R}^{m \times d_1}$ are initialized i.i.d. from $\mathcal{N}(0, 0.02^2)$. We set $d_1 = 100$, $d = 20$, $\alpha_1 = 5$, $\alpha_2 = 0.5$, $m = 50$. Noisy inputs are $x_\tau = x_0 + \tau\epsilon$ with $\tau \in [0, 2.33]$ (larger τ = higher noise).

We compare three protocols[1]: **Standard training.** We sample $\tau \sim \mathrm{Unif}(0, 2.33)$ throughout training, with all neurons active. **High-to-low denoising schedule only.** We start with $\tau \sim \mathrm{Unif}(1.5, 2.33)$ (high-noise only) and progressively decrease $\tau_{\min}$ from 1.5 to 0 across stages, ending

[1]We do not consider model sparsity alone (without denoising scheduling), as it performs significantly worse than standard training in preliminary experiments.

at $\tau \sim \mathrm{Unif}(0, 2.33)$. All neurons remain active. **Joint denoising-sparsity scheduling.** We couple the denoising schedule with high-to-low model sparsity: $\tau_{\min}$ decreases from 1.5 to 0 while neurons are progressively unfrozen, starting from 50% active and ending with all active.

High-to-low training induces sequential feature learning with improved purity. For each feature $\mathbf{M}_{1j}$, we rank neurons w_i by their normalized projection $|\langle w_i, \mathbf{M}_{1j}\rangle|/\|w_i\|$ and average this over the top 20 neurons. A value near 1 indicates a purified representation of $\mathbf{M}_{1j}$, while mixed representations yield smaller projections due to dilution. Figure 3(top) reports this metric averaged over all $\mathbf{M}_1$ features; Figure 3(middle) shows the analogous quantity for $\mathbf{M}_2$. All methods successfully learn $\mathbf{M}_1$-aligned neurons, with joint scheduling yielding slightly purer representations. The major advantage of joint scheduling is on $\mathbf{M}_2$: standard training consistently learns $\mathbf{M}_2$ features mixed with $\mathbf{M}_1$, since $\mathbf{M}_1$ dominates. Denoising scheduling alone improves over standard but still fails to purify $\mathbf{M}_2$. In contrast, joint scheduling keeps a subset of neurons inactive during the high-noise phase, fully avoiding $\mathbf{M}_1$ interference; these neurons activate later and learn $\mathbf{M}_2$ features from low-noise data, achieving significantly higher $\mathbf{M}_2$ alignment.

Silhouette Score reflects the sequential learning of class-discriminative and irrelevant features. We divide the samples into d classes, where the class labels depend only on $\mathbf{M}_1$ and are independent of $\mathbf{M}_2$. Specifically, for each sample we draw z_1 to have exactly one active coordinate $j \in [d]$ and assign the class label of x_0 to be that index j, while z_2 is sampled independently so that $\mathbf{M}_2 z_2$ contributes class-irrelevant variation. For each sample x_0, we compute the hidden representation $h = \sigma(Wx_0)$ and evaluate group separation using the Silhouette Score (SS) with cosine distance (Yu et al., 2023; Mo et al., 2024; Zhang et al., 2022). The SS provides a feature-agnostic measure of how well representations are separated according to class labels: higher SS indicates stronger separation, while lower SS indicates weaker separation.

As shown in Figure 3 (bottom), under joint denoising-sparsity scheduling, the SS score increases rapidly during the early stage of training. This trend coincides with the phase in which neurons predominantly align with $\mathbf{M}_1$ features (Figure 3, top). After approximately 6K iterations, the SS begins to decrease, which temporally aligns with the emergence of alignment with $\mathbf{M}_2$ features (Figure 3, middle). This pattern can be interpreted as follows. Since class labels depend only on $\mathbf{M}_1$, stronger separation (higher SS) suggests that the learned hidden representation h is more aligned with $\mathbf{M}_1$. In later stages, as h additionally incorporates $\mathbf{M}_2$ features, which are not discriminative for the labels, the separation between classes may be reduced, leading to a decrease in SS. Thus, while SS does not directly identify which features are learned, the up-then-down trend of SS score is consistent with a transition from learning predominantly class-discriminative features to incorporating less relevant features.

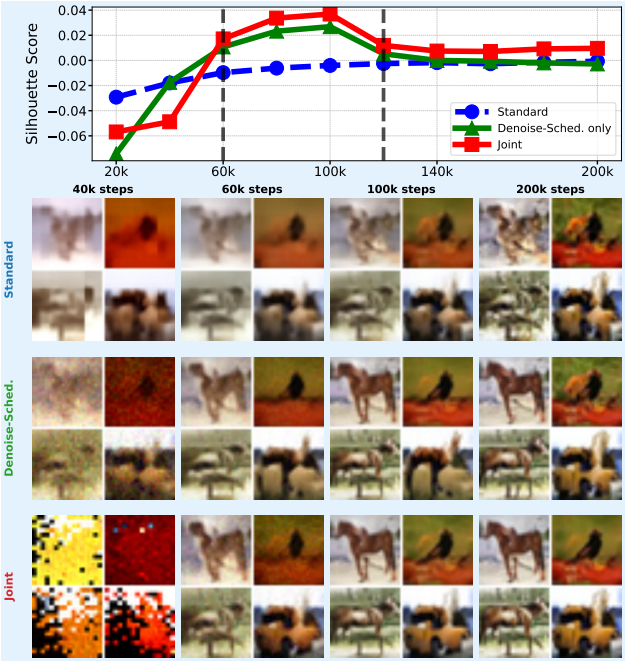

Figure 4. Silhouette Score on CIFAR-10 using U-VIT. High-to-low methods achieve higher SS during early training, then SS drops as fine-grained learning begins which is the same with Fig 3 (bottom). Standard shows stable SS throughout.

5.2. Experiments on Practical Data and Models

We evaluate our method on CIFAR-10 and CelebA-64 using the U-ViT architecture (Bao et al., 2023). We adopt the standard forward process with a linear $\beta(t)$ schedule (Ho et al., 2020). We compare three training protocols that differ only in timestep sampling and model capacity scheduling: **Standard training.** Under the linear $\beta(t)$ schedule, we sample timesteps uniformly from the full range $t \sim \mathrm{Unif}(0, 1)$ throughout training (with smaller t corresponding to lower noise).

High-to-low denoising schedule. We start by sampling $t \sim \mathrm{Unif}(0.5, 1)$ only high-noise data and then progressively decrease $t_{\min}$ from 0.5 to 0 over multiple stages, until reaching $t \sim \mathrm{Unif}(0, 1)$. **Joint denoising-sparsity scheduling.** We couple the high-to-low denoising schedule with high-to-low model sparsity. Across stages, we progressively decrease $t_{\min}$ from 0.5 to 0, and simultaneously expand the patch-embedding capacity by gradually unfreezing more output channels C_{out}[2]. We provide more experimental de-

[2]In U-ViT, the patch embedding is a Conv2d layer with weight shape $(C_{\text{out}}, 3, 4, 4)$. Our framework expands C_{out} from 166 (65%) to 256 (100%) for CIFAR-10 dataset.

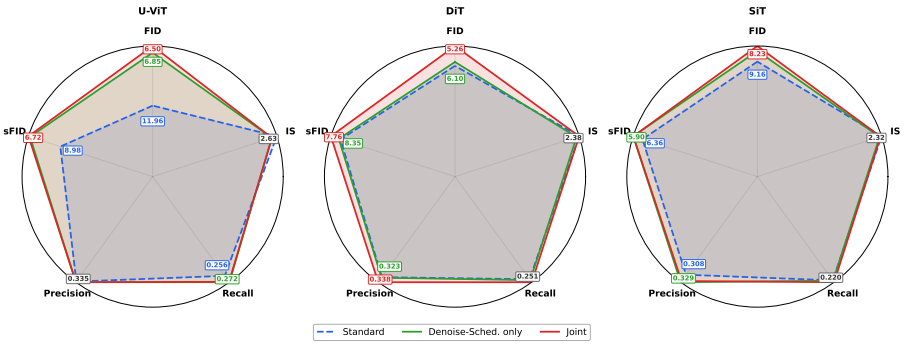

Figure 5. CelebA-64 comparison across three diffusion backbones (U-ViT, DiT, SiT). Each radar shows five metrics at the argmin-FID checkpoint: FID, sFID (↓); Precision, Recall, IS (↑). Radial axis is ratio-to-best per panel; vertex labels show raw values.

tails in Appendix A.1 and CNN-based U-Net (Ho et al., 2020) results in Appendix A.19.

High-to-low training induces staged learning on real data. We extract hidden representations from the middle block of U-ViT, located between the encoder and decoder, and compute the SS score for these representations using the object classes of CIFAR-10 as groups. Figure 4 shows the evolution of the SS score over training. Under joint denoising–sparsity scheduling, the SS score increases rapidly during the early stages and begins to decline after approximately 100K iterations. This up-then-down trend closely mirrors the behavior observed in the synthetic setting (Figure 3, bottom), where the transition coincides with the model first learning the dominant $\mathbf{M}_1$ features and subsequently shifting to $\mathbf{M}_2$ features (Figure 3, top and middle). Although such feature groups cannot be directly identified in practical datasets, the similar trend in the SS score suggests a potentially similar underlying dynamic. We therefore conjecture that the increase phase reflects the learning of class-discriminative patterns, while the subsequent decrease corresponds to the model learning less relevant features. Moreover, high-to-low denoising schedule exhibits a similar trend with smaller changes. In contrast, standard training shows a relatively stable SS trajectory with smaller values. Similar patterns are observed on CelebA, with additional results provided in Appendix A.2.

We also show the generative quality of models obtained at different training iterations in Figure 4 (bottom). At 100K iterations, images generated by joint denoising-sparsity scheduling capture the structural information of the main features. Images from high-to-low denoising schedule show slightly lower quality, and both methods produce much clearer images than the baseline. By the end of training, images from both joint scheduling and denoising schedule contain more fine-grained details, and the quality of these images remain substantially higher than that of the baseline.

Joint denoising-sparsity scheduling achieves best FID and fastest convergence. Figure 2 shows FID vs. training FLOPs and generated samples on CelebA-64 using U-ViT. With the joint scheduling, FID drops below 7 at roughly 730 PFLOPs of training compute, whereas the denoising-schedule-only model needs about 860 PFLOPs ($\sim 18\%$ more compute) to reach the same level. In contrast, standard training is still above FID 14 around 774 PFLOPs and barely reaches 12 at 860 PFLOPs, never matching the converged level of the other two methods. The bottom panel displays generated samples at the same final training step (200k) for all three methods: the joint scheduling consistently produces sharper facial details and fewer distortions at equal training step.

We conduct extensive empirical evaluations of our joint denoising-sparsity scheduling across four diffusion backbones, including U-ViT (Bao et al., 2023), DiT (Peebles & Xie, 2023), SiT (Ma et al., 2024), and U-Net (Ho et al., 2020), on four image datasets: CIFAR-10, CelebA-64, FFHQ-64, and AFHQ-64. For each backbone and dataset pair, we compare against standard training and high-to-low denoising baselines, reporting FID curves throughout training, samples at representative checkpoints, and a multi-metric summary at the argmin-FID checkpoint. As an illustrative example, Figure 5 compares the three training strategies on CelebA-64 across the three transformer-based backbones under five metrics: FID, sFID, Precision, Recall, and IS, where lower FID/sFID and higher Precision/Recall/IS indicate better performance. Each model is evaluated at its argmin-FID checkpoint. Joint denoising-sparsity scheduling consistently achieves the best FID and sFID across all three architectures, while remaining competitive on Precision, Recall, and IS. Detailed setup, FID curves, generated samples, and per-architecture multi-metric summaries are provided in Appendix A.1, A.9, A.14, and A.19.

6. Conclusion

We propose a joint high-to-low denoising and sparsity framework for diffusion models, with theory showing that aligning noise schedules with capacity allocation yields purified feature learning. Training should extend beyond denoising schedules to sparsity, offering both theoretical insight and practical guidance.

Impact Statement

This paper provides theoretical analysis of high-to-low training strategies in diffusion models. Our work is primarily theoretical and aims to improve the understanding of training dynamics in DMs. The proposed joint denoising-sparsity scheduling may help reduce computational costs by improving training efficiency, which could lower the environmental impact of training large-scale DMs. While DMs in general can be misused for generating misleading content, our theoretical contributions do not directly enable such misuse beyond existing capabilities. We do not foresee specific negative societal consequences arising from this work.

Following prior DM theory, our analysis is on two-layer ReLU networks; extending to deeper architectures remains open. We also focus on high-to-low denoising and sparsity scheduling; other strategies such as data augmentation, learning-rate scheduling, and loss weighting are out of scope. On the empirical side, we did not evaluate on large-scale benchmarks such as ImageNet-256, as our study requires comparing multiple training algorithms across many runs and architectures, which is prohibitive at that scale under our compute budget; we leave such large-scale validation to future work.

References

Allen-Zhu, Z. and Li, Y. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 977–988. IEEE, 2022.

Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., and Zhu, J. All are worth words: A vit backbone for diffusion models. In *CVPR*, 2023.

Block, A., Mroueh, Y., and Rakhlin, A. Generative modeling with denoising auto-encoders and langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

Boffi, N. M., Jacot, A., Tu, S., and Ziemann, I. Shallow diffusion networks provably learn hidden low-dimensional structure. *arXiv preprint arXiv:2410.11275*, 2024.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In *ICLR*, 2023.

Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., and Yoon, S. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11472–11481, 2022.

Conforti, G., Durmus, A., and Menickelly, M. Score diffusion models without early stopping: finite fisher information is all you need. *arXiv preprint arXiv:2308.12240*, 2023.

Croitoru, F. A., Hondru, V., Ionescu, R. T., Sebe, N., and Shah, M. Curriculum direct preference optimization for diffusion and consistency models. In *Proceedings of CVPR*, 2025.

Han, A., Huang, W., Cao, Y., and Zou, D. On the feature learning in diffusion models. *arXiv preprint arXiv:2412.01021*, 2024.

Hang, T., Gu, S., Li, C., Bao, J., Chen, D., Hu, H., Geng, X., and Guo, B. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7441–7451, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Hoogeboom, E., Heek, J., and Salimans, T. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pp. 13213–13232. PMLR, 2023.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Kadkhodaie, Z., Guth, F., Simoncelli, E. P., and Mallat, S. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *ICLR*, 2024.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35: 26565–26577, 2022.

Kim, J.-Y., Go, H., Kwon, S., and Kim, H.-G. Denoising task difficulty-based curriculum for training diffusion models. *arXiv preprint arXiv:2403.10348*, 2024.

Lee, H., Lu, J., and Tan, Y. Convergence of score-based generative modeling for general data distributions. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pp. 946–985. PMLR, 2023.

Li, W., Zhang, H., and Qu, Q. Shallow diffuse: Robust and invisible watermarking through low-dimensional subspaces in diffusion models. *arXiv preprint arXiv:2410.21088*, 2024.

Li, X., Zhang, Z., Li, X., Chen, S., Zhu, Z., Wang, P., and Qu, Q. Understanding representation dynamics of diffusion models via low-dimensional modeling. *arXiv preprint arXiv:2502.05743*, 2025.

Liang, Y., Bhardwaj, S., and Zhou, T. Diffusion curriculum: Synthetic-to-real data curriculum via image-guided diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1697–1707, 2025.

Ma, N., Goldstein, M., Albergo, M. S., Boffi, N. M., Vanden-Eijnden, E., and Xie, S. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Mo, Y., Lu, Z., Yu, R., Zhu, X., and Wang, X. Revisiting self-supervised heterogeneous graph learning from spectral clustering perspective. *Advances in Neural Information Processing Systems*, 37:43133–43163, 2024.

Oko, K., Akiyama, S., and Suzuki, T. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pp. 26517–26582. PMLR, 2023.

Pareek, D., Oh, S., and Du, S. S. Understanding the gain from data filtering in multimodal contrastive learning. *arXiv preprint arXiv:2512.14230*, 2025.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *ICCV*, 2023.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Wang, B. and Pehlevan, C. An analytical theory of spectral bias in the learning dynamics of diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Wang, B. and Vastola, J. J. Diffusion models generate images like painters: an analytical theory of outline first, details later. *arXiv preprint arXiv:2303.02490*, 2023.

Wen, Z. and Li, Y. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pp. 11112–11122. PMLR, 2021.

Yu, C., Shi, Y., and Wang, J. Contextually affinitive neighborhood refinery for deep clustering. *Advances in Neural Information Processing Systems*, 36:5778–5790, 2023.

Zhang, X., Zhao, Z., Tsiligkaridis, T., and Zitnik, M. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in neural information processing systems*, 35:3988–4003, 2022.

Overview of Appendix

The overall structure of the appendix is as follows. Each appendix provides supplementary information that supports the main content of this document but is not included in the main body to maintain clarity and flow.

An impact statement and a discussion of limitations are included at the beginning of the appendix after thhis overview.

Appendix A provides additional numerical experiments across four diffusion-model backbones and multiple image datasets. Specifically, we evaluate U-ViT (Bao et al., 2023) on CIFAR-10, CelebA-64 (at $D = 256$ and $D = 512$), and AFHQ-64, including a Silhouette Score analysis on CelebA-64 and a Jacobian eigenvector analysis; DiT (Peebles & Xie, 2023) on CelebA-64, FFHQ-64, and AFHQ-64 ($h = 512$); SiT (Ma et al., 2024) on CelebA-64, FFHQ-64, and AFHQ-64; and a U-Net (Ho et al., 2020) backbone on CIFAR-10 and CelebA-64. For each backbone we report FID curves under our joint denoising-sparsity scheduling against baselines, together with multi-metric summaries across datasets.

Appendices B–J present the proofs of the main theoretical results.

Appendix B introduces the notation and provides a proof sketch. We recommend that readers begin with Appendix B before reading the detailed proofs. Appendix C collects the key technical lemmas and preliminary derivations that support the analysis in the remaining appendices.

Appendices D–I develop the training dynamics and generalization analysis for the three training approaches considered in this paper: standard training, high-to-low noise scheduling only, and joint scheduling. Appendices D–E focus on the early training stage. In these stages, all three methods exhibit the same dynamics for the non-frozen neurons, while the Joint approach additionally includes a subset of frozen neurons that remain unchanged. Appendices F–H study Stage II, where the three methods begin to exhibit different learning behaviors. Appendix I analyzes the convergence behavior after the model has learned the features. Finally, Appendix J builds on these convergence results and establishes the generalization guarantees for the models obtained by each training method.

Specifically, we have

- **Impact Statement and Limitations**
- **Appendix A: Extra Experiment**
  Additional experiments, including Transformer-based diffusion model evaluations and extended analysis.
- **Appendix B: Notations and Proof Sketchy**
  The notation used throughout the appendices and a proof sketch of the main theoretical analysis. The proof sketch outlines the key steps connecting the training-dynamics analysis, convergence results, and generalization guarantees.
- **Appendix C: Technical Lemmas and Preliminary Derivations**
  The supporting lemmas, mathematical preliminaries, and preliminary derivations used throughout the theoretical analysis.
- **Appendix D: Stage 0 – Feature Initialization**
  Analysis of network initialization and geometry of lucky neurons.
- **Appendix E: Stage I – $\mathbf{M}_1$ Alignment Phase**
  Theoretical proof of $\mathbf{M}_1$ feature learning during the first training phase.
- **Appendix F: Stage II – $\mathbf{M}_2$ Alignment (High-to-Low Schedule)**
  Theoretical proof of $\mathbf{M}_2$ feature learning under high-to-low denoising schedule.
- **Appendix G: Stage II – $\mathbf{M}_2$ Alignment (Joint Scheduling)**
  Theoretical proof of $\mathbf{M}_2$ feature learning under joint denoising-sparsity scheduling.
- **Appendix H: Stage II – $\mathbf{M}_2$ Alignment (Standard Training)**
  Analysis of entangled representations when using standard training without scheduling.
- **Appendix I: Convergence of Training Dynamics**
  Proof of convergence for the diffusion training dynamics.

- **Appendix J: Generalization**
  Generalization bounds and analysis for the learned representations.

A. Extra Experiment

All experiments were conducted on an internal compute cluster using NVIDIA H100 GPUs, and each run completed within 50 GPU-hours.

A.1. Transformer-based Diffusion Model Experiments

Model Architecture. We adopt U-ViT (Bao et al., 2023) as the backbone architecture, which is a Vision Transformer (ViT) based diffusion model with long skip connections inspired by U-Net. U-ViT treats all inputs (the time embedding, condition embedding, and noisy image patches) as tokens and processes them with a stack of transformer blocks. The architecture consists of an encoder with $L/2$ transformer blocks, a middle block, and a decoder with $L/2$ transformer blocks. Long skip connections connect corresponding encoder and decoder blocks, enabling effective gradient flow and multi-scale feature fusion.

Each transformer block contains multi-head self-attention (with Flash Attention for efficiency), layer normalization, and an MLP with GELU activation. The time step t is embedded via sinusoidal positional encoding and prepended to the patch sequence as an extra token. The model predicts the noise ϵ added to the input image (noise prediction parameterization).

Training Configuration. Table 1 summarizes the training hyperparameters across the four datasets. All models are trained for 200K iterations with the AdamW optimizer using mixed-precision (FP16) on multiple GPUs in parallel. The four datasets share the same transformer depth, head count, and MLP ratio; they differ in image size, patch size, and embedding dimension. CelebA-64 has two variants ($D = 256$ and $D = 512$) to study the effect of model width. For high-to-low methods, we define multiple stages that progressively expand the noise level range from high noise ($t_{\min} = 0.3$) to full range ($t_{\min} = 0.0$).

Table 1. Training configuration for the four U-ViT experiments. Only per-dataset differences are listed; shared hyperparameters are summarized in the bottom row.

Hyperparameter	**CIFAR-10**	**CelebA-64 (D=256)**	**CelebA-64 (D=512)**	**AFHQ-64**
Image size	32×32	64×64	64×64	64×64
Patch size	2	4	4	4
Embedding dimension D	256	256	512	256
Warmup steps	2,500	5,000	5,000	5,000
Shared across all runs: transformer depth $L = 12$, 8 attention heads, MLP ratio 4, 256 patches, 200,000 total iterations, batch size 128 per GPU, AdamW (lr 2×10^{-4}, weight decay 0.03), FP16 precision; sampling for FID uses Euler-Maruyama SDE with 1,000 steps.				

Training Schedule. For high-to-low methods, we use a multi-stage schedule that progressively expands the noise level range. For joint denoising-sparsity scheduling, we additionally apply a sparsity schedule that gradually unlocks the embedding capacity. The noise level $t \in [t_{\min}, t_{\max}]$ controls which diffusion timesteps are sampled during training, where $t = 1.0$ corresponds to pure noise and $t = 0.0$ corresponds to clean images. Early stages restrict training to high noise levels ($t_{\min} \geq 0.3$) so the model first learns coarse structure; as training progresses $t_{\min}$ decreases to include finer details. For the sparsity schedule (joint scheduling only), we apply an embedding sparsity mask after the patch embedding layer[3]. This forces early-stage learning to use a compact representation, encouraging neurons to focus on the most salient features first; sparsity is then progressively reduced to unlock capacity for fine-grained features.

CIFAR-10, CelebA-64 ($D = 256$), and CelebA-64 ($D = 512$) share the same eight-stage timing in Table 2; the active-dimension column scales with D (e.g. $D = 512$ at sparsity 0.15 exposes $436/512$ dimensions). AFHQ-64 contains roughly $11\times$ fewer images than CelebA, so we shorten the masked-embedding budget and front-load it (Table 3).

All schedules transition to full-range training ($t_{\min} = 0$, sparsity $= 0$) in the final stage, allowing the model to refine all features with full capacity.

[3]In U-ViT, the patch embedding is a `nn.Conv2d` with output shape (B, N, D) where D is the embedding dimension. Our sparsity mask zeros out the last $\lfloor D \times \text{sparsity} \rfloor$ dimensions, effectively reducing the active capacity to $D_{\text{active}} = D - \lfloor D \times \text{sparsity} \rfloor$.

Table 2. Shared training schedule for CIFAR-10 (32×32, patch $= 2$), CelebA-64 (D=256), and CelebA-64 (D=512). Active-dim values shown for $D = 256$; the $D = 512$ run uses $D_{\text{active}} = 512 - \lfloor 512 \, \text{sparsity} \rfloor$ at the same sparsity values.

Stage	$t_{\min}$	$t_{\max}$	Steps	Cumulative	Sparsity	Active Dims (D=256)
Joint High-to-Low (Denoise + Sparsity)						
1	0.30	1.0	10K	10K	0.15	218/256
2	0.20	1.0	10K	20K	0.15	218/256
3	0.10	1.0	20K	40K	0.15	218/256
4	0.07	1.0	10K	50K	0.13	223/256
5	0.05	1.0	10K	60K	0.10	231/256
6	0.03	1.0	20K	80K	0.08	236/256
7	0.01	1.0	20K	100K	0.05	244/256
8	0.00	1.0	100K	200K	0.00	256/256

Table 3. Training schedule for AFHQ-64 (64×64, patch $= 4$, $D = 256$). The high-noise stages run longer and the masked-embedding budget is compressed to 80K (vs. 100K elsewhere), reflecting the smaller dataset size ($\sim$14.6K images).

Stage	$t_{\min}$	$t_{\max}$	Steps	Cumulative	Sparsity	Active Dims
Joint High-to-Low (Denoise + Sparsity)						
1	0.30	1.0	20K	20K	0.15	218/256
2	0.20	1.0	15K	35K	0.15	218/256
3	0.10	1.0	12K	47K	0.15	218/256
4	0.07	1.0	10K	57K	0.13	223/256
5	0.05	1.0	8K	65K	0.10	231/256
6	0.03	1.0	8K	73K	0.08	236/256
7	0.01	1.0	7K	80K	0.05	244/256
8	0.00	1.0	120K	200K	0.00	256/256

A.2. Silhouette Score Analysis on CelebA-64

Joint denoising-sparsity scheduling learns purer feature representations on CelebA-64. To verify our theoretical prediction that high-to-low training induces purer feature representations, we compute the silhouette score of intermediate features during training. Specifically, we extract hidden representations from the middle block of U-ViT, located between the encoder and decoder (output shape: $B \times 257 \times 256$, where the first token is the time embedding and the remaining 256 tokens are patch embeddings). Following the CIFAR-10 analysis in Fig 4, we use facial attributes as grouping criteria. We select four representative attributes: gender (Female vs Male), bangs (No Bangs vs Bangs), heavy makeup (Light vs Heavy), and hair color (Black, Blond, Brown, Gray). The reported silhouette score is the average across these four attributes.

Figure 6 shows that both high-to-low methods exhibit the same up-then-down trend observed in synthetic data (Figure 3) and CIFAR-10 (Figure 4): silhouette scores peak during intermediate training, then decrease as the model captures finer-grained features. In contrast, standard training maintains consistently lower scores throughout, confirming entangled representations as predicted by Theorem 4. The generated samples corroborate this: high-to-low methods produce recognizable faces earlier while baseline outputs remain blurry.

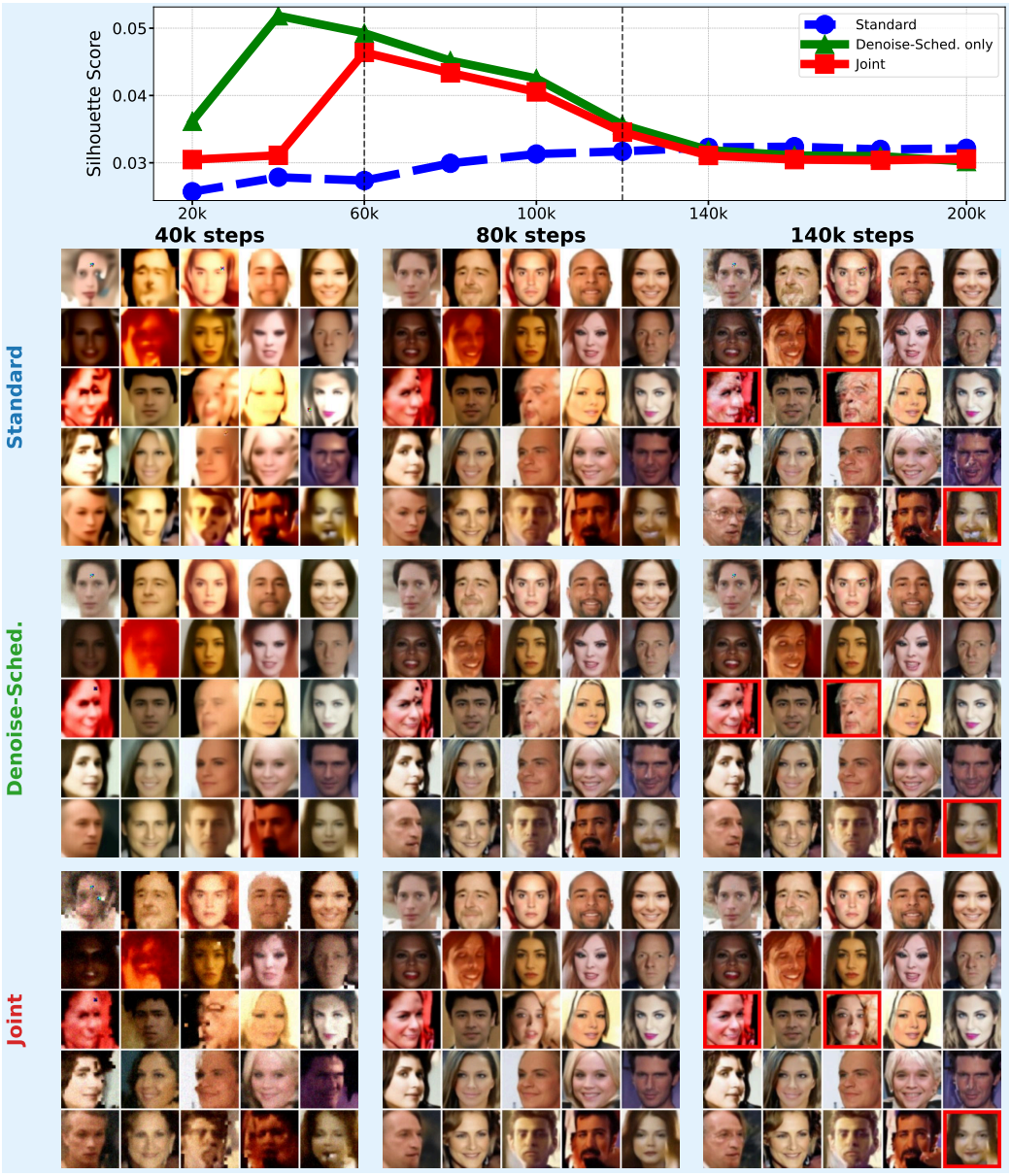

Figure 6. Feature purity comparison on CelebA-64. **Top:** Silhouette score (cosine distance) computed from the middle block of U-ViT, averaged over four facial attributes (gender, bangs, heavy_makeup, hair_color). Joint scheduling peaks around 60K iterations, then decreases as the model learns finer-grained features. Vertical dashed lines indicate phase transitions at 60K and 120K iterations. **Bottom:** Generated samples at 40K, 80K, and 140K training steps.

A.3. FID Analysis on CIFAR-10

Figure 7 shows FID curves and generated samples on CIFAR-10 (D=256). At 200K steps, joint scheduling reaches FID 10.81, ahead of denoising schedule (11.44) and standard training (11.61). Joint also converges fastest: it crosses each FID milestone before the other two methods. Red boxes highlight three representative samples (a bird, a truck, and a car). At 200K steps, joint scheduling produces sharper object boundaries and more accurate category-specific features; denoising schedule exhibits slight category ambiguity (e.g., the car shows blurred boundaries resembling a ship), while standard training shows the most artifacts.

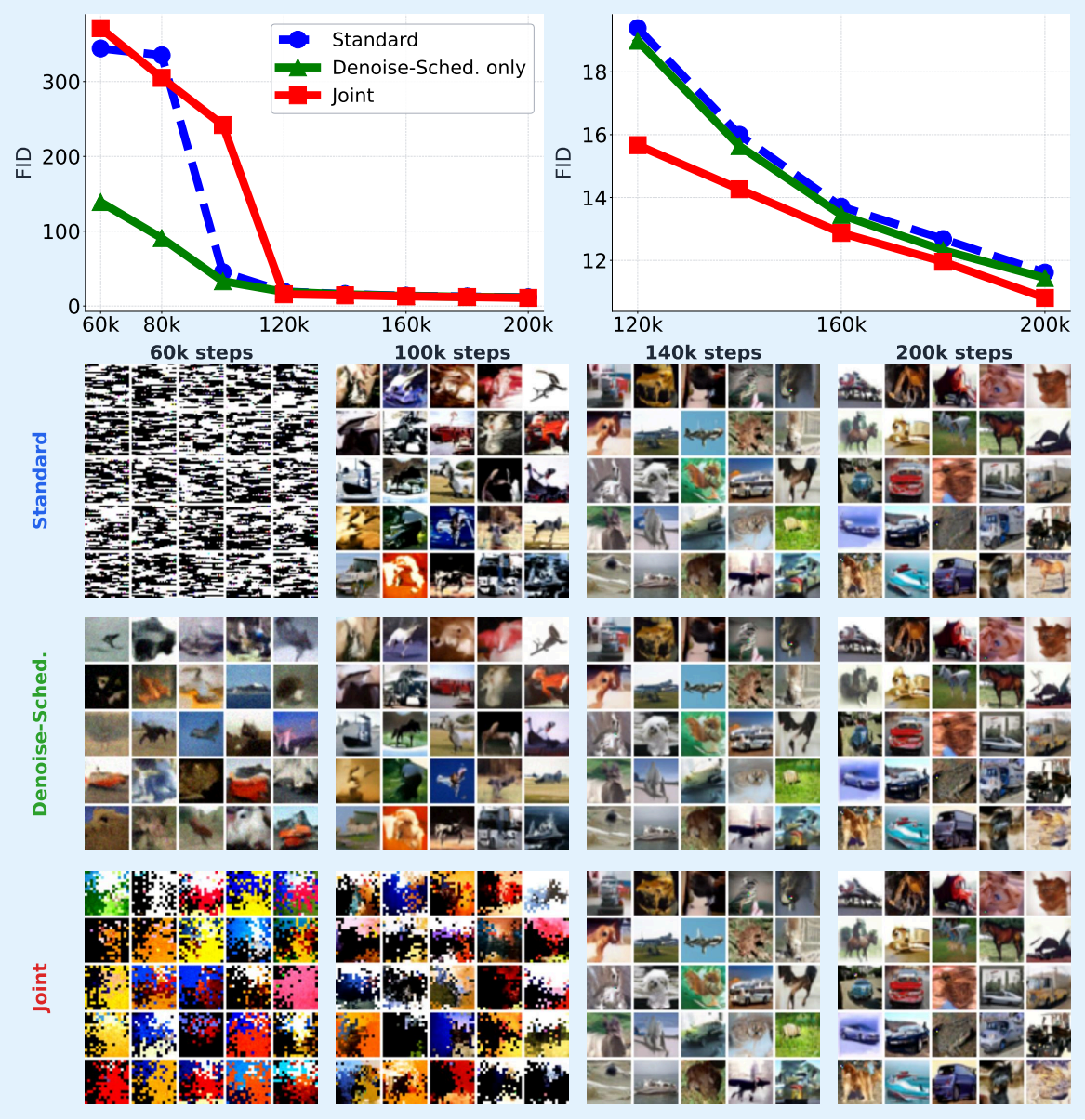

Figure 7. **CIFAR-10 (D=256) generation quality comparison. Top:** FID curves from 80K to 200K steps. Left panel shows overall convergence; right panel zooms in on 100K to 200K. **Bottom:** Sample images at 60K, 80K, 100K, and 200K steps. Red boxes highlight samples where joint scheduling shows clearest improvements in object clarity and category accuracy.

A.4. FID Analysis on CelebA-64 ($D = 256$)

We evaluate our training strategies on CelebA-64 with embedding dimension $D = 256$. Figure 8 shows the FID curves and representative samples. Joint scheduling reaches FID 6.50 at 200K, well below denoising schedule (6.85) and standard

training (11.96). The gap is largest in early training: joint reaches FID $\sim$8 around 150K, where standard training is still above 17. Joint produces sharper eye details, better-defined facial contours, and more natural skin textures, while standard training exhibits noticeable artifacts and blurring at the same step.

Figure 8. FID curves and generated samples on CelebA-64 (D=256). **Top:** FID from 140K to 200K. **Bottom:** Generated samples at 140K, 160K, 180K, and 200K. Joint scheduling produces visibly cleaner faces well before standard training catches up.

A.5. FID Analysis on CelebA-64 ($D = 512$)

The wider variant ($D = 512$) is included to verify that the benefits of joint scheduling persist at increased model width. Figure 9 reports FID curves and samples. At 200K, joint reaches FID 4.07, ahead of standard training (4.21) and denoising schedule (4.82). Although the absolute gap narrows compared to $D = 256$ (the wider model fits CelebA-64 well even without scheduling), joint still converges fastest and achieves the lowest final FID.

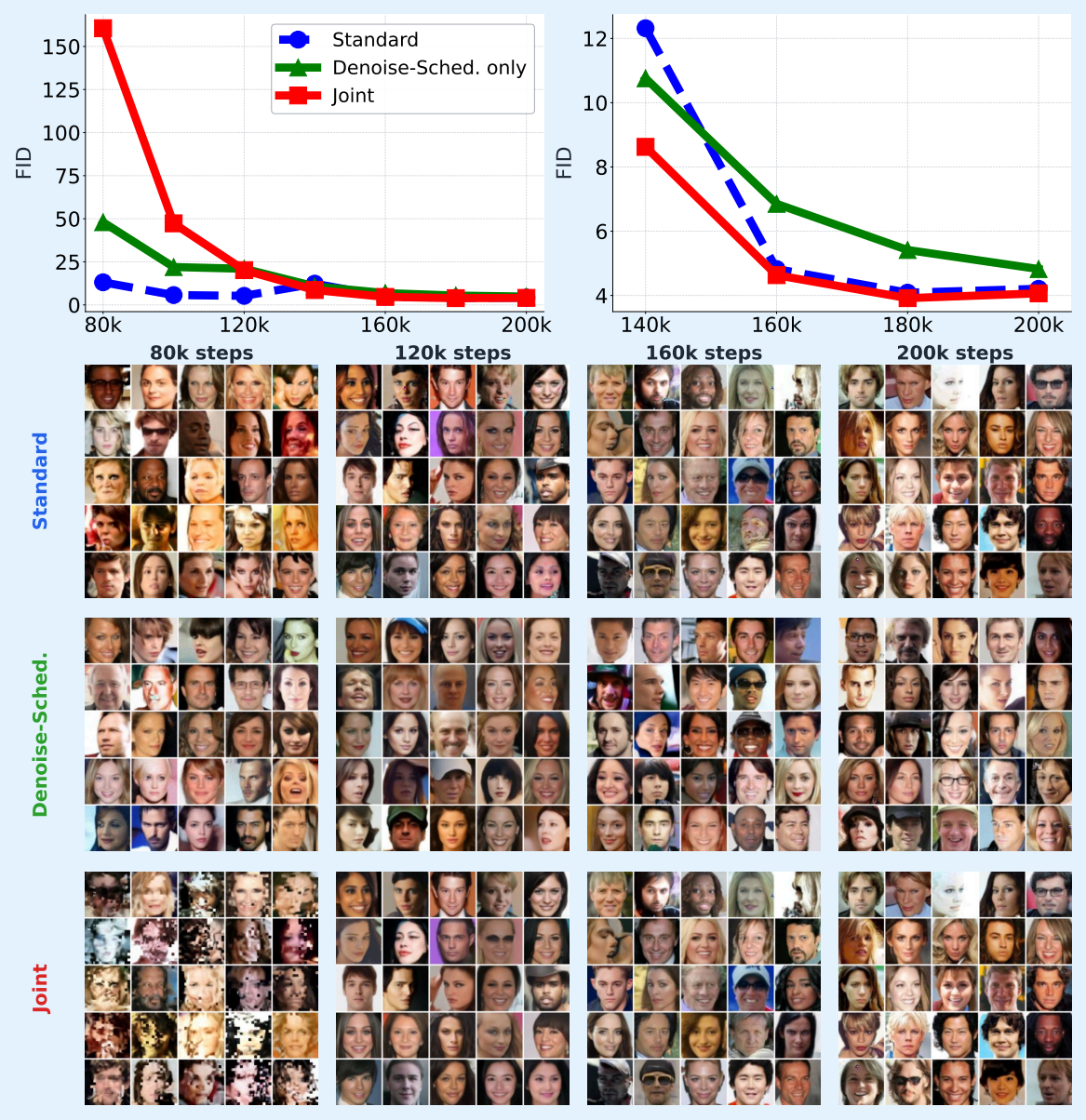

Figure 9. FID curves and generated samples on CelebA-64 (D=512). The wider U-ViT closes most of the gap that standard training left on $D = 256$; joint scheduling still leads.

A.6. FID Analysis on AFHQ-64

AFHQ-64 contains roughly 14.6K images (about 11× smaller than CelebA-64). With less data, the masked-embedding stages must be shorter to avoid over-training under reduced capacity, motivating the front-loaded schedule in Table 3. Figure 10 shows the result: joint scheduling reaches FID 9.75 at 200K, ahead of denoising schedule (11.13) and standard training (12.59). Joint samples show cleaner fur textures and fewer high-frequency artifacts, especially in cat faces and dog ears.

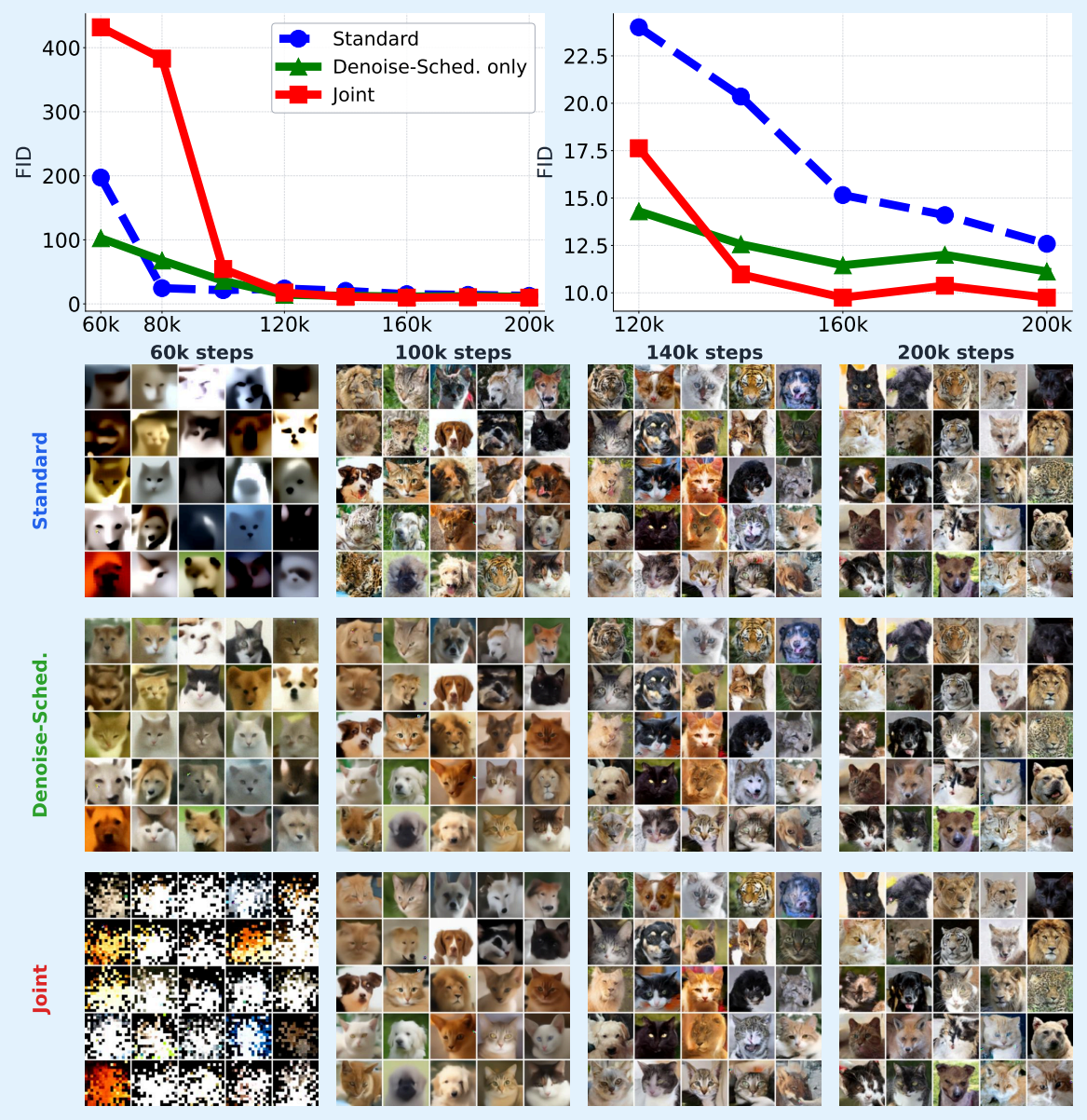

Figure 10. FID curves and generated samples on AFHQ-64 (D=256). Joint scheduling holds the lead throughout training and produces the cleanest fine-grained details at 200K.

A.7. Multi-Metric Summary across Datasets

Figure 11 consolidates the multi-metric comparison across all four U-ViT datasets at each model's argmin-FID checkpoint, evaluated with the same protocol used in Figure **??** (FID/sFID lower better, Precision/Recall/IS higher better; radial coordinate is ratio-to-best within the panel, vertex labels are raw values). Joint scheduling holds the best FID and sFID

on every dataset and stays at least competitive on Precision/Recall/IS, confirming that the FID gains in the curves above translate into broader generation-quality improvements rather than trading off on other axes.

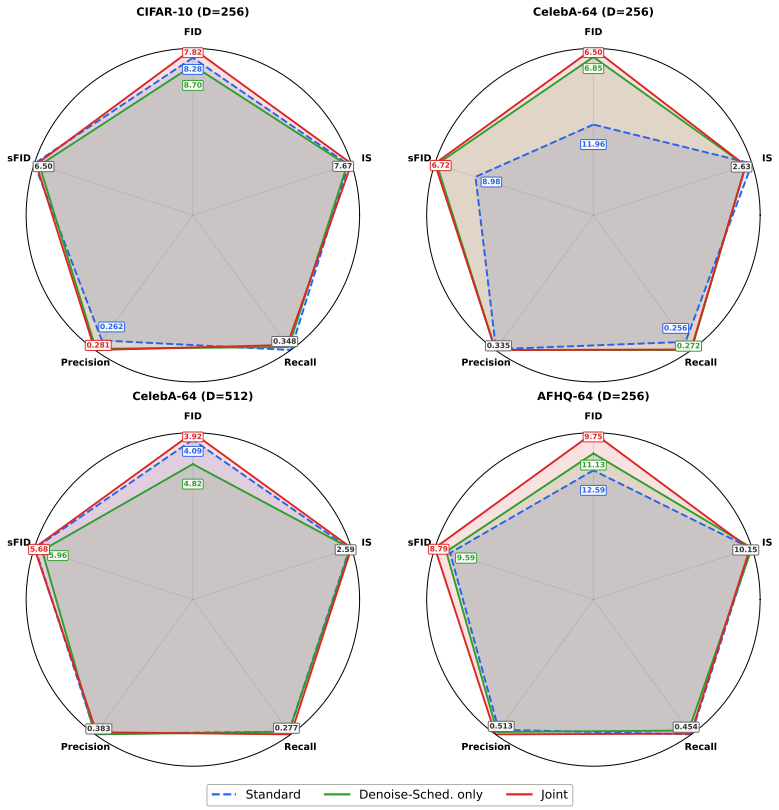

Figure 11. Multi-metric summary across the four U-ViT datasets. Each panel reports five metrics at the argmin-FID checkpoint; radial coordinate is ratio-to-best within the panel, and vertex labels are raw values. Joint scheduling (red) leads on FID and sFID across all four datasets while remaining close to the best on the higher-is-better axes.

A.8. Jacobian Eigenvector Analysis

Following (Kadkhodaie et al., 2024), we analyze the learned denoiser by examining its Jacobian matrix. For a denoiser $f : \mathbb{R}^n \to \mathbb{R}^n$, we compute the Jacobian $J = \nabla f(x)$ and perform SVD on $(I - J) = U\Sigma V^T$. The eigenvectors $\{u_k\}$ reveal the basis in which the denoiser operates, while the eigenvalues $\{\lambda_k\}$ indicate the shrinkage applied along each direction. The leading eigenvectors with larger eigenvalues capture low-frequency semantic structures (analogous to $\mathbf{M}_1$), while eigenvectors with smaller eigenvalues correspond to high-frequency fine details (analogous to $\mathbf{M}_2$).

Figure 12 shows that Joint Scheduling preserves 6 semantic eigenvectors (λ_1-λ_6), while Standard training preserves only 3 (λ_1-λ_3). This result provides two key insights: (1) it validates our multi-scale sparse coding model, as the learned denoisers indeed exhibit a spectral separation between coarse and fine features; (2) it confirms that Joint Scheduling learns richer semantic representations, capturing more $\mathbf{M}_1$ and $\mathbf{M}_2$ directions compared to standard training.

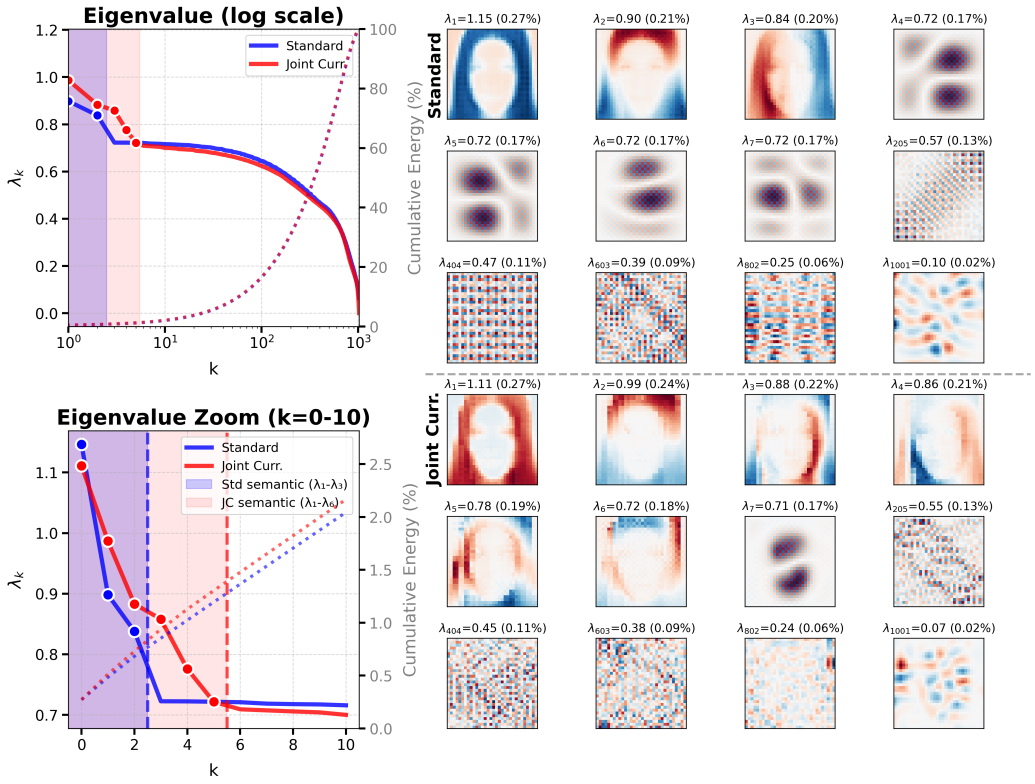

Figure 12. Jacobian eigenvector analysis of $(I - J)$ comparing Standard and Joint Scheduling training. **Top-left:** Eigenvalue spectrum λ_k with cumulative energy ratio. **Bottom-left:** Zoomed view for $k = 0$-10: Standard preserves 3 semantic eigenvectors (λ_1-λ_3), Joint Scheduling preserves 6 (λ_1-λ_6). **Top-right:** Eigenvector visualizations for Standard training. **Bottom-right:** Eigenvector visualizations for Joint Scheduling.

A.9. DiT-based Diffusion Model Experiments

Model Architecture. We use DiT (Peebles & Xie, 2023), a diffusion transformer in which timestep and (optional) class conditioning enter every block via adaLN-Zero modulation rather than as input tokens. Each block stacks self-attention, an MLP with GELU, and learned scale/shift/gate vectors predicted from the conditioning embedding. We train at 64×64 resolution with patch size 4 (256 tokens). For CelebA-64 and FFHQ-64 we use **DiT-S/4** ($D = 384$, 6 heads); for AFHQ-64 we widen to **DiT-S/4 with** $h = 512$ (8 heads), which handles the small ($\sim$14.6K image) AFHQ training set better. The model predicts the noise ϵ (no learned Σ).

Training Configuration. Table 4 lists hyperparameters across the three DiT runs. All runs train for 100K iterations with AdamW, FP16, and Euler-Maruyama SDE sampling at 1,000 steps for FID evaluation.

Table 4. Training configuration for the three DiT-S/4 experiments.

Hyperparameter	**CelebA-64**	**FFHQ-64**	**AFHQ-64**
Image size	64×64	64×64	64×64
Patch size	4	4	4
Hidden size D	384	384	512
Attention heads	6	6	8

Shared across all runs: transformer depth $L = 12$, MLP ratio 4, 256 patches, 100,000 total iterations, batch size 128 per GPU, AdamW (lr 2×10^{-4}, weight decay 0.03, $\beta = (0.99, 0.99)$), 2,500 warmup steps, FP16 precision; sampling for FID uses Euler-Maruyama SDE with 1,000 steps.

Training Schedule. DiT trains $2\times$ faster than U-ViT in our setup, so we compress the U-ViT 200K ladder 1:2 into 100K total iterations: 35K of high-to-low curriculum followed by 65K of full-range training. CelebA-64 and FFHQ-64 share the

standard ladder in Table 5; AFHQ-64 ($h = 512$) uses a gentler entry (Table 6) because the smaller dataset combined with the wider model is more sensitive to early masked-embedding stages.

Table 5. Shared DiT-S/4 training schedule (CelebA-64 and FFHQ-64). Active dimensions are computed for $D = 384$ as $D - \lfloor D \cdot \text{sparsity} \rfloor$.

Stage	$t_{\min}$	$t_{\max}$	Steps	Cumulative	Sparsity	Active Dims
Joint High-to-Low (Denoise + Sparsity)						
1	0.30	1.0	5K	5K	0.20	308/384
2	0.20	1.0	5K	10K	0.15	327/384
3	0.10	1.0	5K	15K	0.15	327/384
4	0.07	1.0	5K	20K	0.13	335/384
5	0.05	1.0	5K	25K	0.10	346/384
6	0.03	1.0	5K	30K	0.08	354/384
7	0.01	1.0	5K	35K	0.05	365/384
8	0.00	1.0	65K	100K	0.00	384/384
High-to-Low Denoising Schedule Only						
1	0.30	1.0	5K	5K	–	–
2	0.20	1.0	5K	10K	–	–
3	0.10	1.0	5K	15K	–	–
4	0.07	1.0	5K	20K	–	–
5	0.05	1.0	5K	25K	–	–
6	0.03	1.0	5K	30K	–	–
7	0.01	1.0	5K	35K	–	–
8	0.00	1.0	65K	100K	–	–

Table 6. DiT-S/4 ($h = 512$) training schedule on AFHQ-64. Stage 1 is extended to 10K at 0.15 sparsity (gentler entry than the 5K $\times$ 0.20 used elsewhere) and stage 8 is compressed to 60K to keep the total at 100K. Active dimensions use $D = 512$.

Stage	$t_{\min}$	$t_{\max}$	Steps	Cumulative	Sparsity	Active Dims
Joint High-to-Low (Denoise + Sparsity)						
1	0.30	1.0	10K	10K	0.15	436/512
2	0.20	1.0	5K	15K	0.15	436/512
3	0.10	1.0	5K	20K	0.15	436/512
4	0.07	1.0	5K	25K	0.13	446/512
5	0.05	1.0	5K	30K	0.10	461/512
6	0.03	1.0	5K	35K	0.08	472/512
7	0.01	1.0	5K	40K	0.05	487/512
8	0.00	1.0	60K	100K	0.00	512/512
High-to-Low Denoising Schedule Only						
1	0.30	1.0	10K	10K	–	–
2	0.20	1.0	5K	15K	–	–
3	0.10	1.0	5K	20K	–	–
4	0.07	1.0	5K	25K	–	–
5	0.05	1.0	5K	30K	–	–
6	0.03	1.0	5K	35K	–	–
7	0.01	1.0	5K	40K	–	–
8	0.00	1.0	60K	100K	–	–

A.10. FID Analysis on CelebA-64 with DiT

Figure 13 shows FID curves and samples for DiT-S/4 on CelebA-64. At 100K steps (n=50K SDE-1000), joint scheduling reaches FID 5.26, ahead of denoising schedule (5.99) and standard training (6.20). Joint converges fastest throughout the 100K window, and the bottom panel shows joint samples retain cleaner facial details earlier than the other two methods.

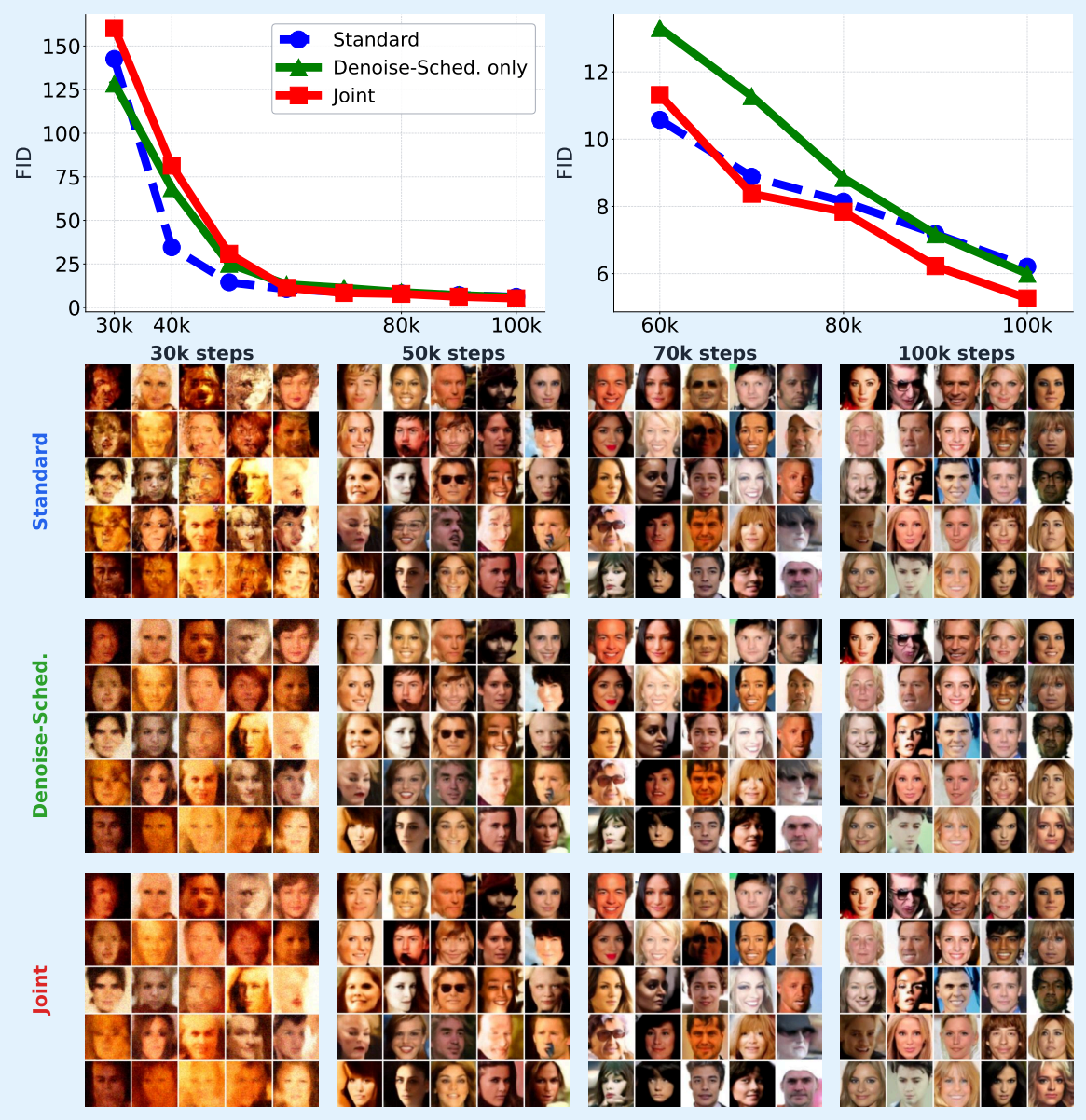

Figure 13. FID curves and generated samples on CelebA-64 with DiT-S/4. Joint scheduling produces the lowest final FID and the cleanest faces at each checkpoint.

A.11. FID Analysis on FFHQ-64 with DiT

Figure 14 shows the FFHQ-64 result for DiT-S/4. Joint scheduling reaches FID 8.45 at 100K, well below denoising schedule (9.82) and standard training (10.63). The improvement is consistent across the curve and is also visible in the samples: joint outputs show fewer texture artifacts and more coherent facial structure.

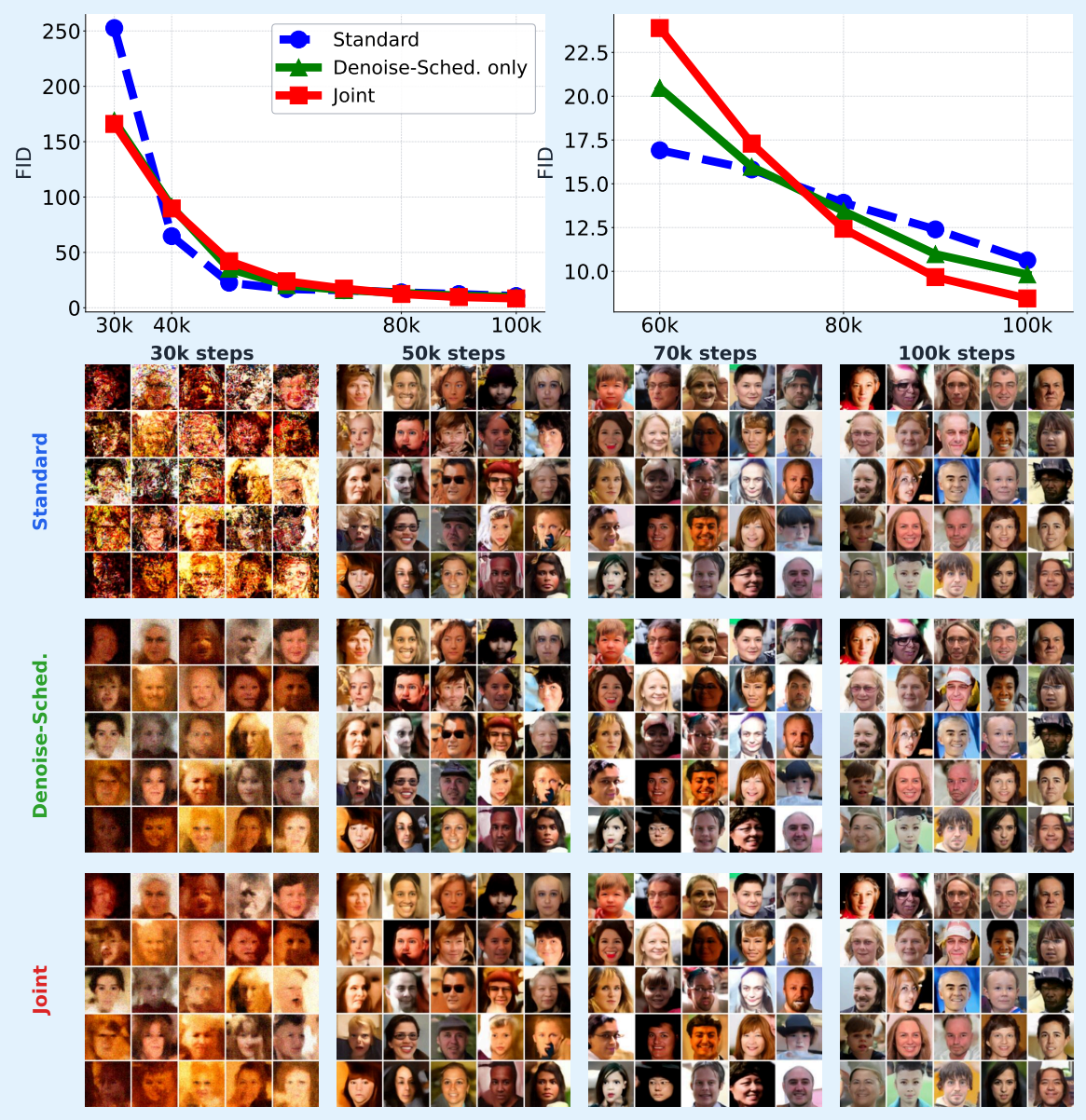

Figure 14. FID curves and generated samples on FFHQ-64 with DiT-S/4. Joint scheduling holds the lowest FID throughout training and produces visibly cleaner samples at 100K.

A.12. FID Analysis on AFHQ-64 with DiT ($h = 512$)

AFHQ-64 has only ∼14.6K training images, so we run DiT at $h = 512$ to give the wider model enough capacity for fine fur and feather textures. Figure 15 reports the result. The three methods land within ∼0.25 FID of one another (Standard 7.85, Denoise 7.88, Joint 8.10); the gentler curriculum entry (Table 6) prevents joint scheduling from over-restricting capacity, but absolute FID is so close that no method clearly wins on this dataset. The radar in Section A.13 shows joint still leads on Precision while trailing slightly on FID.

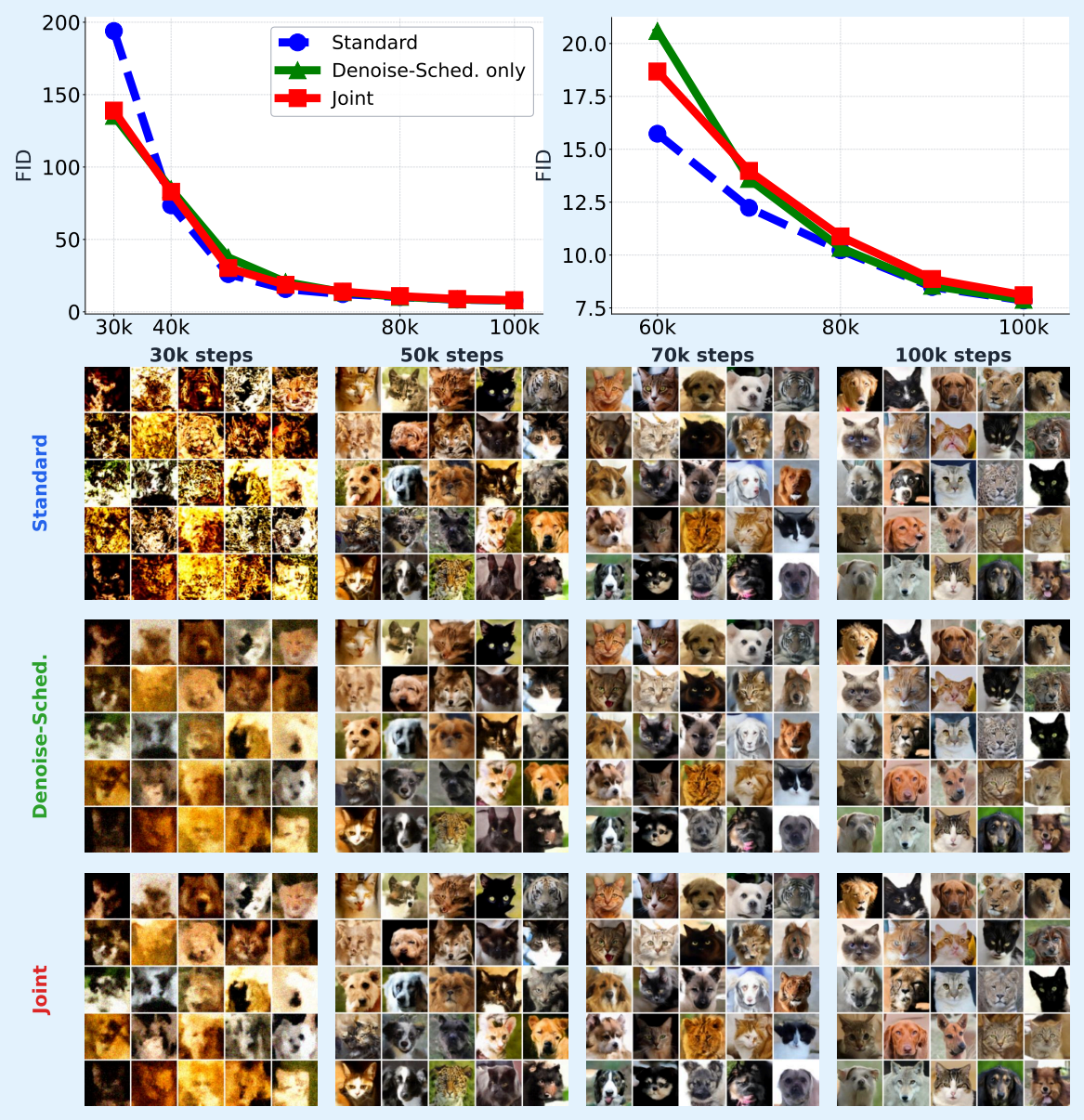

Figure 15. FID curves and generated samples on AFHQ-64 with DiT-S/4 ($h = 512$). All three methods converge to a similar FID; differences are small and method-dependent across other metrics (see radar summary).

A.13. Multi-Metric Summary across DiT Datasets

Figure 16 consolidates the multi-metric comparison across the three DiT datasets at each model's argmin-FID checkpoint, evaluated under the same protocol used in Figure **??** (FID/sFID lower better, Precision/Recall/IS higher better; radial coordinate is ratio-to-best within the panel, vertex labels are raw values). Joint scheduling leads on FID and sFID for the two larger datasets (CelebA-64, FFHQ-64) and remains best on Precision for AFHQ-64 ($h = 512$), where the three methods cluster within a narrow FID band.

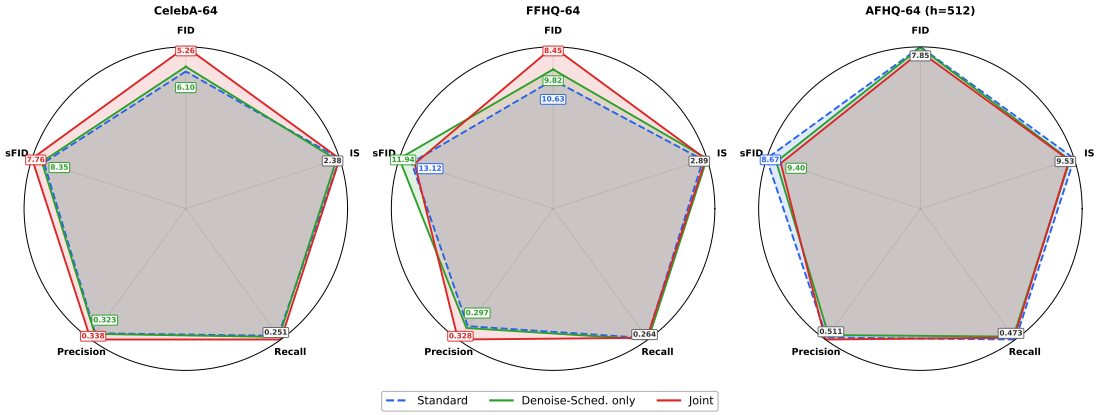

Figure 16. Multi-metric summary across the three DiT-S/4 datasets at each model's argmin-FID checkpoint. Joint scheduling leads on FID/sFID on CelebA-64 and FFHQ-64; on AFHQ-64 ($h = 512$) the methods are tightly clustered with joint best on Precision.

A.14. SiT-based Diffusion Model Experiments

Model Architecture. We use SiT (Ma et al., 2024), an interpolant-based diffusion transformer that re-uses the DiT block layout but replaces the noise-prediction objective with a velocity / interpolant-flow objective. We instantiate **SiT-T/4** ($D = 256$, 8 heads, depth $L = 12$, MLP ratio 4) at 64×64 resolution with patch size 4 (256 tokens). FID is evaluated with Euler-Maruyama *ODE* sampling at 1,000 steps (the SiT default), unlike the SDE-1000 sampling used for DiT and U-ViT.

Training Configuration. Table 7 lists hyperparameters across the three SiT runs. All three datasets use the same SiT-T/4 backbone; only the training data and dataloader change.

Table 7. Training configuration for the three SiT-T/4 experiments.

Hyperparameter	**CelebA-64**	**AFHQ-64**	**FFHQ-64**
Image size	64×64	64×64	64×64
Patch size	4	4	4
Hidden size D	256	256	256
Attention heads	8	8	8
Shared across all runs: transformer depth $L = 12$, MLP ratio 4, 256 patches, 100,000 total iterations, batch size 128 per GPU, AdamW (lr 2×10^{-4}, weight decay 0.03, $\beta = (0.99, 0.99)$), 2,500 warmup steps, FP16 precision; sampling for FID uses Euler-Maruyama ODE with 1,000 steps.			

Training Schedule. The three SiT runs share the same 8-stage curriculum used by the DiT CelebA / FFHQ runs (35K curriculum + 65K full-range = 100K), shown in Table 8. Active dimensions match those of the U-ViT $D = 256$ schedule.

Table 8. Shared SiT-T/4 training schedule across CelebA-64, AFHQ-64, and FFHQ-64.

Stage	$t_{\min}$	$t_{\max}$	Steps	Cumulative	Sparsity	Active Dims
Joint High-to-Low (Denoise + Sparsity)						
1	0.30	1.0	5K	5K	0.20	205/256
2	0.20	1.0	5K	10K	0.15	218/256
3	0.10	1.0	5K	15K	0.15	218/256
4	0.07	1.0	5K	20K	0.13	223/256
5	0.05	1.0	5K	25K	0.10	231/256
6	0.03	1.0	5K	30K	0.08	236/256
7	0.01	1.0	5K	35K	0.05	244/256
8	0.00	1.0	65K	100K	0.00	256/256
High-to-Low Denoising Schedule Only						
1	0.30	1.0	5K	5K	–	–
2	0.20	1.0	5K	10K	–	–
3	0.10	1.0	5K	15K	–	–
4	0.07	1.0	5K	20K	–	–
5	0.05	1.0	5K	25K	–	–
6	0.03	1.0	5K	30K	–	–
7	0.01	1.0	5K	35K	–	–
8	0.00	1.0	65K	100K	–	–

A.15. FID Analysis on CelebA-64 with SiT

Figure 17 shows FID curves and samples for SiT-T/4 on CelebA-64. At 100K steps (n=50K ODE-1000), joint scheduling reaches FID 8.07, ahead of denoising schedule (8.38) and standard training (9.16). Joint maintains the lead from very early in training and the samples show progressively cleaner facial structure compared to the other two.

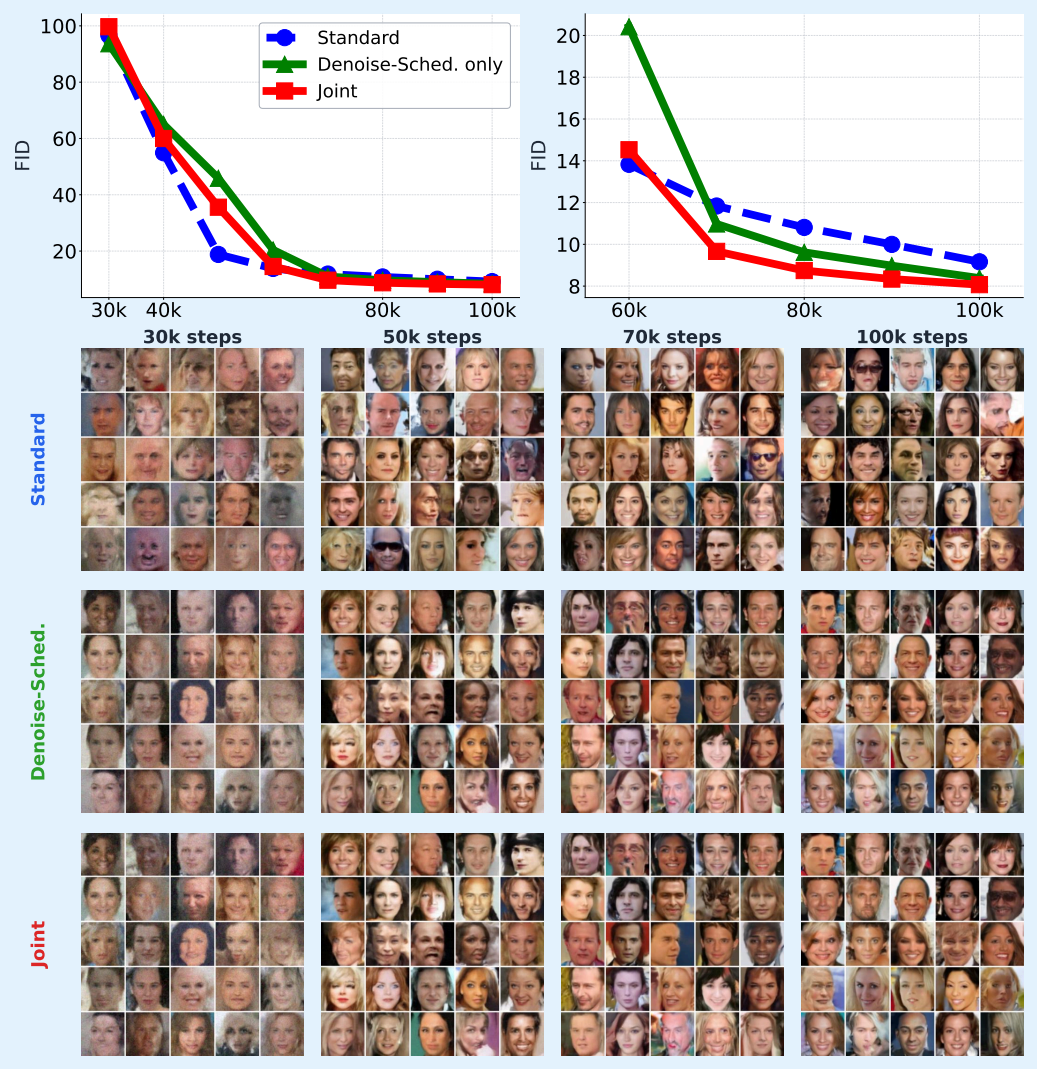

Figure 17. FID curves and generated samples on CelebA-64 with SiT-T/4. Joint scheduling achieves the lowest FID and produces sharper facial details than the baselines.

A.16. FID Analysis on AFHQ-64 with SiT

Figure 18 reports SiT-T/4 on AFHQ-64. Both denoising schedule (14.72) and joint scheduling (14.74) outperform standard training (17.23) by roughly 2.5 FID points; the gap between denoise-only and joint is within run-to-run noise on this small ($\sim$14.6K) dataset. Joint still wins on sFID (6.13 vs. 6.55 vs. 6.84), indicating better local feature statistics even when global FID ties.

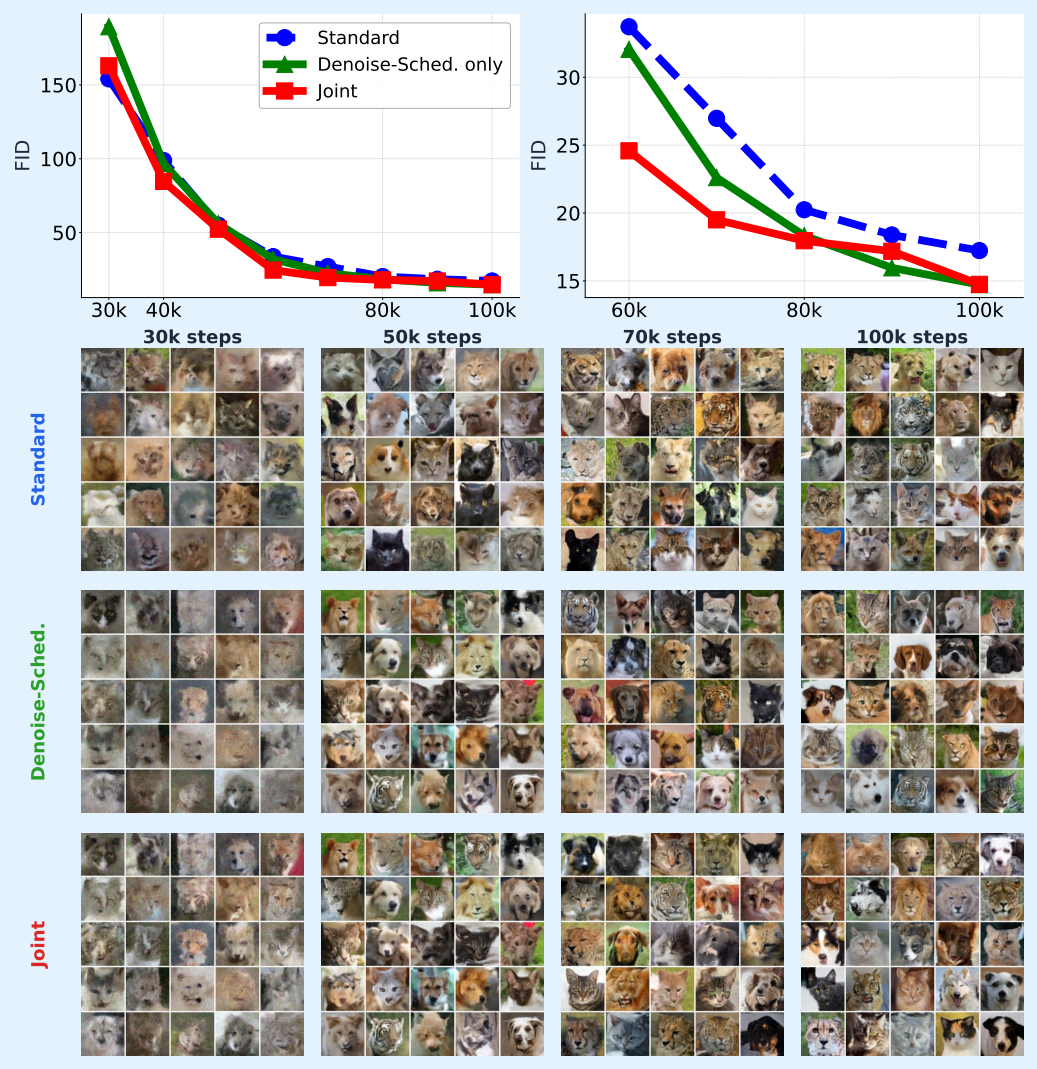

Figure 18. FID curves and generated samples on AFHQ-64 with SiT-T/4. Both high-to-low methods clearly improve over standard; joint matches denoising schedule on FID and edges ahead on sFID and Precision.

A.17. FID Analysis on FFHQ-64 with SiT

Figure 19 shows the FFHQ-64 result for SiT-T/4. Joint scheduling reaches FID 14.44, ahead of denoising schedule (14.77) and standard training (16.70). Joint also leads on Precision and ties denoise-only on Recall, consistent with the radar summary in Section A.18.

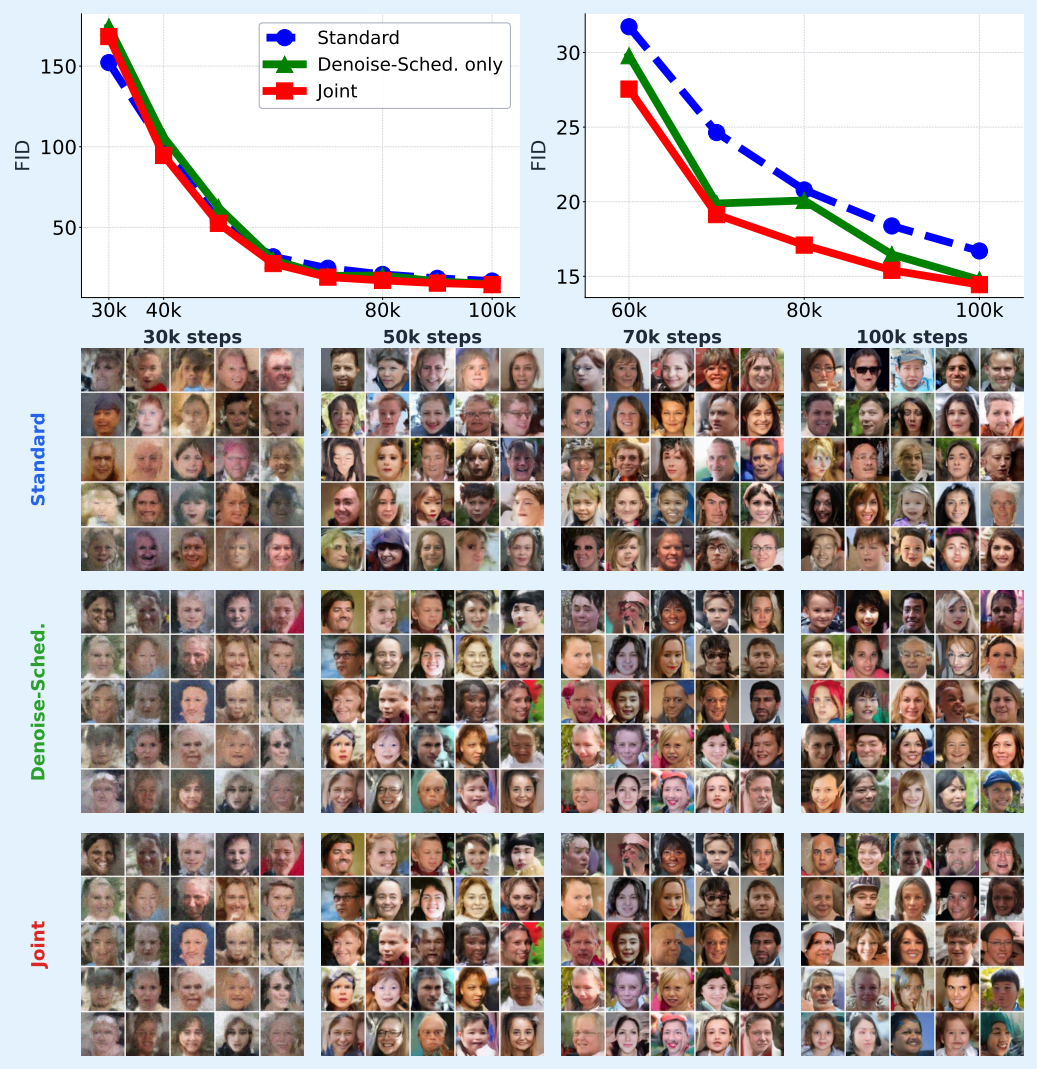

Figure 19. FID curves and generated samples on FFHQ-64 with SiT-T/4. Joint scheduling produces the lowest FID and shows fewer texture artifacts than standard training.

A.18. Multi-Metric Summary across SiT Datasets

Figure 20 consolidates the multi-metric comparison across the three SiT-T/4 datasets at each model's argmin-FID checkpoint. Joint scheduling leads on FID for CelebA-64 and FFHQ-64, ties denoising schedule on FID for AFHQ-64, and is the best or tied-best on sFID and Precision across all three datasets.

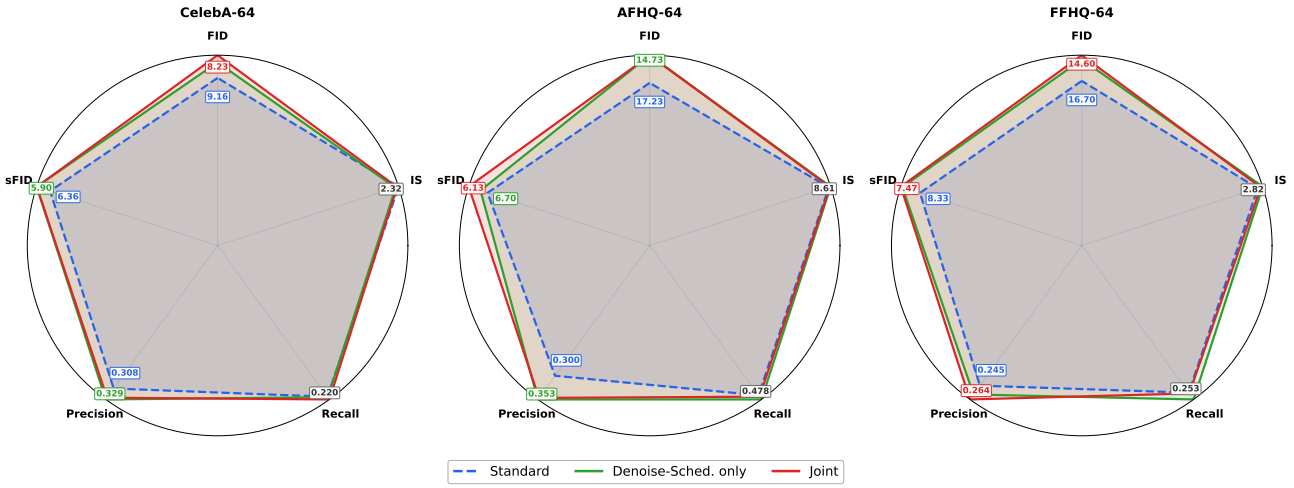

Figure 20. Multi-metric summary across the three SiT-T/4 datasets at each model's argmin-FID checkpoint. Joint scheduling leads on FID and sFID on CelebA-64 and FFHQ-64; on AFHQ-64 it ties denoising schedule on FID and edges ahead on sFID and Recall.

A.19. U-Net-based Diffusion Model Experiments

Model Architecture. We adopt the U-Net architecture from Ho et al. (2020) as the backbone for our experiments. The model follows an encoder-decoder structure with skip connections between corresponding resolution levels. Each level consists of residual blocks with GroupNorm normalization and SiLU activation. Self-attention is applied at the 16×16 resolution level for both datasets. The timestep t is embedded via sinusoidal positional encoding, projected through a two-layer MLP, and injected into each residual block. The model predicts the noise ϵ added to the input image (noise prediction parameterization).

Training Configuration. Table 9 summarizes the training hyperparameters for CIFAR-10 and CelebA-64 experiments. All models are trained with the Adam optimizer using mixed-precision (FP16) training on 6 GPUs in parallel. For high-to-low methods, we define multiple stages that progressively expand the noise level range from high noise ($t_{\min} = 0.3$) to full range ($t_{\min} = 0.0$).

Table 9. Training configuration for U-Net experiments on CIFAR-10 and CelebA-64.

Hyperparameter	**CIFAR-10**	**CelebA-64**
Model Architecture		
Image size	32×32	64×64
Hidden channels	128	64
Channel multipliers	[1, 2, 2, 2]	[1, 2, 2, 4]
Residual blocks per level	2	2
Attention resolution	16×16	16×16
Bottleneck channels	256	256
Diffusion Process		
Timesteps T	1000	1000
β schedule	Linear	Linear
β_{start}, β_{end}	0.0001, 0.02	0.0001, 0.02
Training		
Total epochs	51	23
Batch size	128	128
Learning rate	2×10^{-4}	2×10^{-5}
Optimizer	Adam	Adam
EMA decay	0.9999	0.9999

Training Schedule. For high-to-low methods, we use a multi-stage schedule that progressively expands the noise level range. For joint scheduling, we additionally apply a sparsity schedule that gradually increases the model capacity. Table 10 shows the detailed training stages for CIFAR-10 and CelebA-64, respectively.

The sparsity schedule is applied at the bottleneck layer of U-Net. For CIFAR-10, the bottleneck has $C_{\text{bottleneck}} = 2 \times 128 = 256$ channels; for CelebA-64, it has $C_{\text{bottleneck}} = 4 \times 64 = 256$ channels. At sparsity s, we mask $(s \times C_{\text{bottleneck}})$ channels by zeroing their activations, leaving only $(1 - s) \times C_{\text{bottleneck}}$ channels active. For example, at sparsity $s = 0.2$, 205 out of 256 bottleneck channels are active. As training progresses, we gradually reduce sparsity to unlock more channels for fine-grained feature learning.

Table 10. Training schedule for U-Net (bottleneck channels = 256).

Stage	$t_{\min}$	$t_{\max}$	**Epochs**	**Cumulative**	**Sparsity**	**Active Ch.**
Joint High-to-Low on CIFAR-10						
1–3	0.30→0.10	1.0	2,2,4	8	0.20	205/256
4–5	0.07→0.05	1.0	2,2	12	0.10	230/256
6–8	0.03→0.00	1.0	2,2,35	51	0.00	256/256
Joint High-to-Low on CelebA-64						
1–3	0.30→0.10	1.0	1,1,2	4	0.20	205/256
4–5	0.07→0.05	1.0	1,1	6	0.10	230/256
6–8	0.03→0.00	1.0	1,1,15	23	0.00	256/256
High-to-Low Denoising Schedule Only (same $t_{\min}$ schedule, no sparsity)						

The FID trends are consistent with the U-ViT experiments: joint scheduling converges faster while achieving competitive final FID.

A.20. FID Analysis on CelebA-64 with U-Net.

Figure 21 shows FID curves and sample visualization on CelebA-64 using the U-Net architecture. The FID trends are consistent with the U-ViT experiments: joint scheduling converges faster while achieving competitive final FID. Both high-to-low methods outperform standard training throughout the training process. At epoch 22, joint scheduling achieves the lowest FID, followed by denoising schedule, while standard training shows the highest FID. The generated samples (bottom rows) demonstrate that high-to-low methods produce clearer facial features with fewer artifacts. Red boxes highlight representative samples where high-to-low methods show superior fine-grained generation: sharper eye details, better-defined facial contours, and more natural skin textures compared to standard training at the same epoch.

A.21. FID Analysis on CIFAR-10 with U-Net.

Figure 22 shows FID curves and sample visualization on CIFAR-10 using the U-Net architecture. The FID trends are consistent with the CelebA-64 experiments: both high-to-low methods significantly outperform standard training throughout the training process. Joint scheduling converges faster in early epochs and achieves competitive final FID with denoising schedule. The right panel zooms into epochs 30–50, where joint scheduling and denoising schedule both reach FID around 28–30, while standard training remains above 29. At epoch 50, high-to-low methods produce sharper object boundaries and more recognizable category-specific features, particularly visible in animal classes (birds, cats, dogs) where fine-grained details matter.

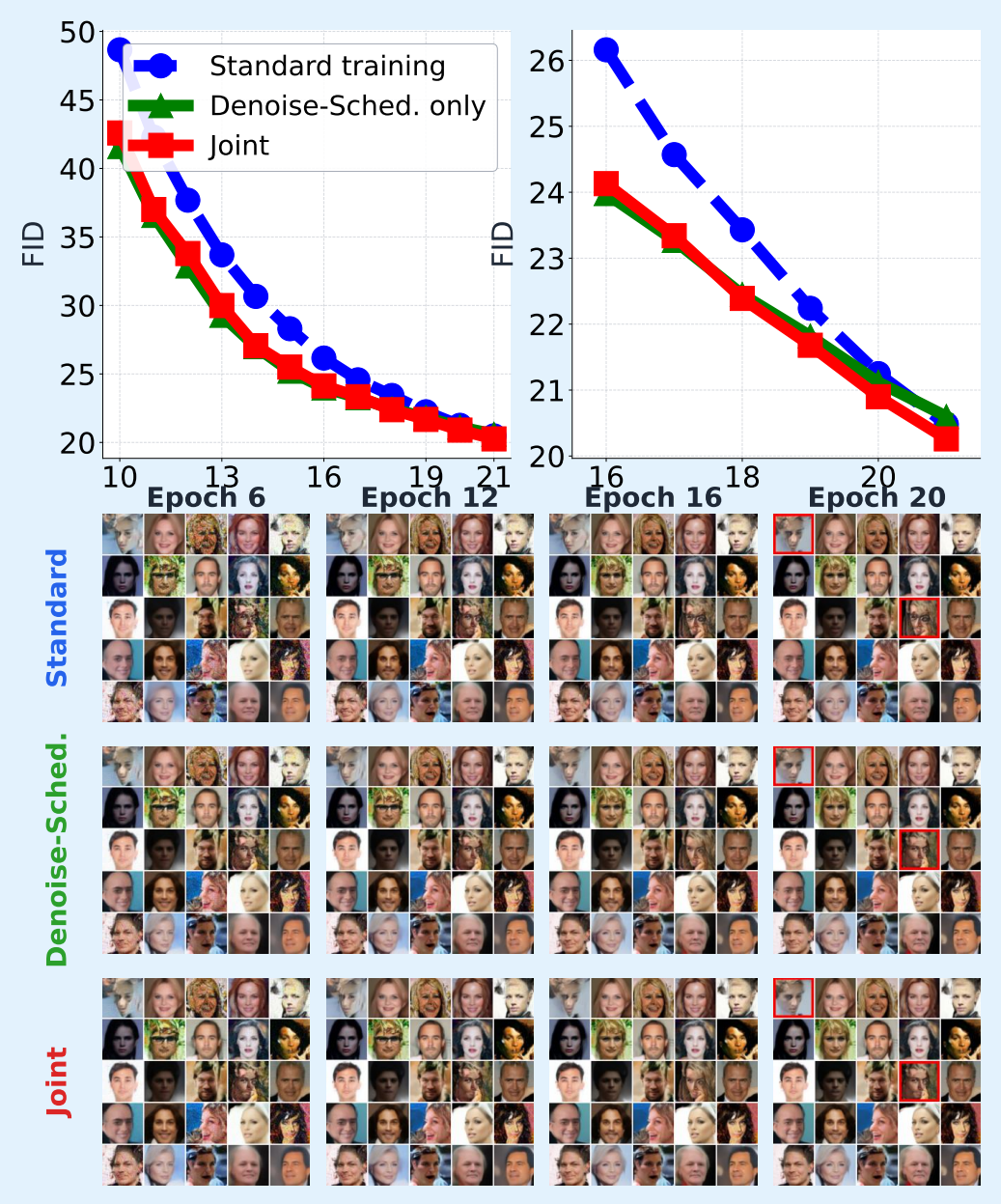

Figure 21. U-Net generation quality comparison on CelebA-64. **Top:** FID curves during training. Left panel shows overall convergence from epoch 10-23; right panel provides detailed view from epoch 14-23. Joint scheduling achieves the best final FID. **Bottom:** Generated samples.

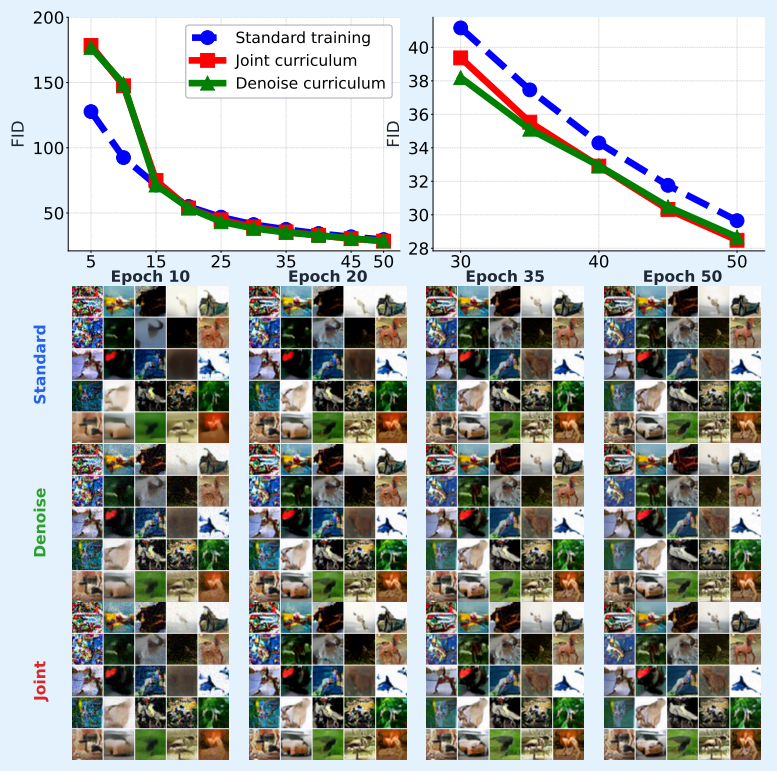

Figure 22. U-Net generation quality comparison on CIFAR-10. **Top:** FID curves during training. Left panel shows overall convergence from epoch 5–50; right panel provides detailed view from epoch 30–50. Both high-to-low methods outperform standard training, with joint scheduling showing fastest early convergence. **Bottom:** Generated samples at epochs 10/25 (left) and 35/50 (right). Red boxes highlight samples where high-to-low methods show clearer object structures compared to standard training.

B. Notation and Proof Sketch

B.1. Notaions

For ease of reference, we summarize the key notation used throughout the appendix. Page 1 collects the architecture and data-model symbols; Page 2 collects training-dynamics, lucky-neuron, and asymptotic notation.

Symbol	Description
Dimensions and indices	
d_1	ambient input/output dimension
d	dictionary size (number of feature directions per dictionary), $d \le d_1$
m	number of hidden neurons
$i \in [m]$	neuron index
$j \in [d]$	dictionary-column index
$k \in \{1, 2\}$	dictionary index (coarse vs. fine)
Network parameters	
$W = [w_1, \dots, w_m]^\top \in \mathbb{R}^{m \times d_1}$	input weight matrix; w_i is the i-th row
$V = [v_1, \dots, v_m]^\top \in \mathbb{R}^{m \times d_1}$	output weight matrix; v_i is the i-th row
$b = (b_1, \dots, b_m)^\top \in \mathbb{R}^m$	per-neuron ReLU bias / activation threshold
$\sigma(\cdot) = \max(\cdot, 0)$	ReLU activation, applied element-wise
$g(x) = V^\top \sigma(Wx - b)$	one-hidden-layer denoiser
σ_0	init. scale: $w_i^{(0)}, v_i^{(0)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_0^2 I_{d_1})$
η	SGD learning rate
B	SGD mini-batch size
Data model	
$x_0 \in \mathbb{R}^{d_1}$	clean signal: $x_0 = \alpha_1 \mathbf{M}_1 z_1 + \alpha_2 \mathbf{M}_2 z_2$
$\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{d_1 \times d}$	orthonormal dictionaries with $\mathbf{M}_1^\top \mathbf{M}_2 = 0$
$\mathbf{M}_{kj}$	j-th column of $\mathbf{M}_k$ (a feature direction)
$\mathbf{M}^\perp$	orthonormal basis of the orthogonal complement of $\mathrm{span}(\mathbf{M}_1, \mathbf{M}_2)$
$z_1, z_2 \in \{-1, 0, 1\}^d$	sparse codes for coarse / fine features
$\alpha_1 = \Theta(1)$	coarse-feature amplitude
$\alpha_2 = \Theta(1/d^{c_0})$	fine-feature amplitude, with constant $c_0 \in (0, 1)$
$\epsilon \sim \mathcal{N}(0, I_{d_1})$	isotropic Gaussian noise vector
$\xi \sim \mathcal{N}(0, \sigma_\xi^2 I_{d_1})$	dense projection-noise term
τ	noise level (scalar), $\tau \sim \mathrm{Unif}[\tau_{\min}, \tau_{\max}]$
$\tau_{\min}, \tau_{\max}$	lower / upper bound of the noise schedule
$\tau_{\text{high}}, \tau_1$	curriculum threshold separating Stage I and Stage II
$x_\tau = x_0 + \tau\epsilon$	noisy observation
L_{DM}	diffusion training loss, $\frac{1}{2}\mathbb{E}\|g(x_\tau) - x_0\|_2^2$
$g^*(x_\tau) = \mathbb{E}[x_0 \mid x_\tau]$	Bayes-optimal denoiser

(Notation table, continued.)

Symbol	Description
Training-stage timestamps	
T_0	end of Stage 0 (initial warm-up phase)
T_1	end of Stage I ($\mathbf{M}_1$-alignment phase)
T_2	end of Stage II ($\mathbf{M}_2$-alignment phase)
t	SGD iteration index; $w_i^{(t)}, v_i^{(t)}, b_i^{(t)}$ denote iterates at step t
Per-direction effective rates	
$a_1 = \Theta(\eta/d)$	effective alignment rate for $\mathbf{M}_1$-directions
$a_2 = \Theta(\eta/d^{1+c_0})$	effective alignment rate for $\mathbf{M}_2$-directions (suppressed by α_2)
Lucky-neuron sets	
$S_{kj,\mathrm{sure}}$	"sure" lucky set: neurons whose initial $\mathbf{M}_{kj}$-alignment satisfies $\big(\frac{\langle w_i^{(0)},\mathbf{M}_{kj}\rangle+\langle v_i^{(0)},\mathbf{M}_{kj}\rangle}{2}\big)^2 \geq \frac{(c_1+c_2)\log d}{d} \cdot \frac{\|\mathbf{M}_k\mathbf{M}_k^\top w_i^{(0)}\|_2^2+\|\mathbf{M}_k\mathbf{M}_k^\top v_i^{(0)}\|_2^2}{4}$
$S_{kj,\mathrm{pot}}$	"potential" lucky set: same form with the looser threshold $(c_1-c_2)\log d/d$
$S_{j,\mathrm{sure}}^{(t)}$	time-t sure set; n in its definition denotes the number of neurons aggregated
S_1	subset of neurons trained in Stage I of joint scheduling (others frozen)
$\mathcal{N}_i$	set of feature indices entangled by neuron i
$\mathbf{r}_i, \mathbf{s}_i$	residual components of $\bar{w}_i, \bar{v}_i$ orthogonal to the aligned directions
Err_t	cumulative higher-order error term in alignment recursion
Scaling constants	
$c_0 \in (0,1)$	exponent controlling the scale gap $\alpha_2 = \Theta(1/d^{c_0})$
c_1, c_2	threshold constants in the lucky-set definitions; $c_1 + c_2 > 2(1 + c_0 - \gamma c_0)$
$\gamma \in (0,1)$	gap parameter separating sure-set from non-potential-set
Asymptotic notation	
$f = O(g)$	$\exists\, C, n_0$: $\lvert f(n)\rvert \leq C\lvert g(n)\rvert$ for all $n \geq n_0$
$f = \Omega(g)$	$g = O(f)$
$f = \Theta(g)$	$f = O(g)$ and $f = \Omega(g)$
$f = o(g)$	$\lim_{n\to\infty} f(n)/g(n) = 0$
$f = \omega(g)$	$\lim_{n\to\infty} f(n)/g(n) = \infty$ (asymptotic strict dominance)
$\tilde{O}(\cdot), \tilde{\Theta}(\cdot)$	hide poly-logarithmic factors in d
"w.h.p."	with probability at least $1 - e^{-\Omega(d_1)}$

Conventions. All vectors are column vectors unless transposed. Inner products $\langle\cdot,\cdot\rangle$ are Euclidean on $\mathbb{R}^{d_1}$; norms $\|\cdot\|_2$ are Euclidean (ℓ_2). The indicator $\mathbf{1}\{\cdot\}$ is 1 when the condition holds and 0 otherwise; in gradient expressions it arises from differentiating the ReLU. Iteration superscripts $w_i^{(t)}$ refer to the value of w_i *after* t SGD updates, with $w_i^{(0)}$ the initialization.

B.2. Proof Sketch

Before presenting the technical lemmas and the four-stage proof of the main theorems, we provide a high-level roadmap that summarizes the key intuitions and the logical flow of the argument. The full proof spans Stage 0 (warm-up), Stage I ($\mathbf{M}_1$-alignment), Stage II ($\mathbf{M}_2$-alignment), and the final convergence and generalization analysis. This sketch highlights the load-bearing ideas in each stage so that a reader may follow the structure without working through every estimate.

Overall strategy. Our analysis tracks the joint evolution of three quantities for every neuron i and every dictionary direction $\mathbf{M}_{kj}$: (i) the signal alignment $\langle w_i^{(t)}, \mathbf{M}_{kj}\rangle$ and its V-side counterpart $\langle v_i^{(t)}, \mathbf{M}_{kj}\rangle$; (ii) the noise projection $\langle w_i^{(t)}, \xi\rangle$ along the orthogonal noise subspace; and (iii) the ReLU bias $b_i^{(t)}$, which acts as an adaptive activation threshold. The proof establishes that, under the curriculum, signal alignment along the active dictionary grows strictly faster than the bias, while noise projections and irrelevant-direction alignments remain dominated by the bias throughout. Once the signal alignment of a neuron crosses its bias, the ReLU indicator stays open on relevant samples, and the alignment grows multiplicatively

until it saturates at the $\Theta(1)$ scale.

Step 1: Linearization of the alignment recursion. Projecting the SGD update (86) onto $\mathbf{M}_{kj}$ and taking expectation over (x_0, τ, ϵ) yields a coupled two-variable recursion for $(\langle w_i^{(t)}, \mathbf{M}_{kj}\rangle, \langle v_i^{(t)}, \mathbf{M}_{kj}\rangle)$, whose leading-order 2×2 system has the form

$$\begin{bmatrix} 1 & a_k \\ a_k & 1 \end{bmatrix}, \qquad a_1 = \Theta(\eta/d), \quad a_2 = \Theta(\eta/d^{1+c_0}).$$

The dominant eigenvalue $1 + a_k$ governs the alignment growth rate along $\mathbf{M}_{kj}$, and the corresponding eigenvector $(1, 1)$ captures the symmetric coupling between w_i and v_i. This recursion underlies every alignment bound in the proof, with non-leading contributions absorbed into a controlled error term Err_t.

Step 2: Stage 0 — separating sure neurons from non-potential neurons. The Gaussian initialization $w_i^{(0)}, v_i^{(0)} \sim \mathcal{N}(0, \sigma_0^2 I)$ induces a random alignment landscape. Standard Gaussian small-ball and anti-concentration estimates yield a partition of the neurons:

- the sure set $S_{1j,\mathrm{sure}}$ contains neurons whose initial $\mathbf{M}_{1j}$-alignment exceeds the threshold $\sqrt{(c_1 + c_2)\log d/d} \cdot \|\mathbf{M}_1\mathbf{M}_1^\top w_i^{(0)}\|_2$, and which therefore amplify under $(1 + a_1)^{T_0}$ to the constant scale by $t = T_0$;
- the complement $S_{1j,\mathrm{pot}}^c$ contains neurons whose alignment lies below the looser threshold $\sqrt{(c_1 - c_2)\log d/d}$ and which provably remain sub-threshold throughout Stage I.

A union bound over $j \in [d]$ ensures that every direction $\mathbf{M}_{1j}$ admits $\Omega(1)$ sure neurons, so coverage of the dictionary is guaranteed.

Step 3: Stage I — $\mathbf{M}_1$-alignment outpaces the bias. The bias schedule satisfies the multiplicative update

$$b_i^{(t+1)} = \big(1 + \eta/d\big)\, b_i^{(t)},$$

while the $\mathbf{M}_1$-alignment of any $i \in S_{1j,\mathrm{sure}}$ grows at rate $1 + a_1 = 1 + \Theta(\eta/d)$ but starting from a $\sqrt{\log d}$-amplified baseline. The crossing inequality

$$(1 + a_1)^{T_0} \cdot \tfrac{\sqrt{\log d}}{\sqrt{d}} \;\gg\; (1 + \eta/d)^{T_0} \cdot b_i^{(0)}$$

guarantees that, by $t = T_0$, the signal alignment crosses the bias. Once the indicator $\mathbf{1}\{w_i^\top x_\tau \geq b_i\}$ stays activated on $\mathbf{M}_{1j}$-aligned samples, the recursion enters its multiplicative regime and saturates at $\Theta(1)$ alignment by $T_1 = \Theta(d \log d/\eta)$.

Step 4: Stage I — $\mathbf{M}_2$-alignment remains dormant. For directions in $\mathbf{M}_2$, the suppressed rate $a_2 = \Theta(\eta/d^{1+c_0})$ is asymptotically smaller than the bias growth η/d. Even under optimistic activation, the multiplicative drift satisfies

$$\frac{1 + \Theta(\eta/d^{1+c_0})}{1 + \eta/d} = 1 - \Theta(\eta/d) + o(\eta/d) < 1,$$

so $\mathbf{M}_2$-alignment cannot cross the rising bias during Stage I. The noise projection is similarly suppressed via the Noise Projection Bound (Lemma 1), giving $|\langle w_i^{(t)}, \xi\rangle| \leq O(\|w_i\|_2/d^{(1+c_0)/2})$.

Step 5: Stage II — $\mathbf{M}_2$-alignment under the curriculum. At $t = T_1$, the curriculum switches to $\tau \in [\tau_{\min}, \tau_{\mathrm{high}}]$, removing the noise that previously suppressed the $\mathbf{M}_2$ gradient signal, and freezes the $\mathbf{M}_1$-aligned subset S_1. The reduced effective noise level multiplies the $\mathbf{M}_2$-alignment rate to $1 + \Theta(\eta/d)$, matching the bias rate. The increased effective signal $\alpha_2/\tau_{\mathrm{high}} = \Theta(1)$, in contrast to $\alpha_2/\tau_{\max} = \Theta(1/d^{c_0})$, tilts the race in favor of alignment. The same sure-set and first-crossing argument as in Stage I then yields pure $\mathbf{M}_2$-alignment by $T_2 = \Theta(d^{1+2c_0} \log d/\eta)$.

Step 6: Failure modes (standard and denoise-only training). Without the sparsity component:

- Under standard training, τ is sampled uniformly throughout, and so $\mathbf{M}_2$-targeted neurons receive interfering gradient contributions from multiple directions $\mathbf{M}_{2j'}$, producing entanglement of size $|\mathcal{N}_i| \geq \Omega(\sqrt{d}/\log d)$.
- Under the denoise-only curriculum, τ decreases over time but all neurons remain unfrozen, and so $\mathbf{M}_1$-aligned neurons re-enter the $\mathbf{M}_2$ phase and contaminate alignment with a residual $\bar{r}_i \perp \mathrm{span}(\mathbf{M}_2)$ of energy comparable to that of the signal.

Both failure modes are quantified through an analogue of the Stage II recursion with an additional interference or residual term that does not vanish in the limit.

Step 7: From feature alignment to generalization. Pure alignment endows the trained denoiser with a matched-filter structure: for τ in the high-noise regime, both curriculum and standard training recover the $\mathbf{M}_1$ component near-optimally; in the low-noise regime, only the curriculum-trained denoiser recovers $\mathbf{M}_2$, whereas the standard denoiser incurs a constant error along the $\mathbf{M}_2$ subspace. The generalization gap stated in Theorem 5 then follows by combining the alignment bounds of Stages I and II with the bias-variance decomposition of the squared-error risk against the Bayes-optimal denoiser $g^*(x_\tau) = \mathbb{E}[x_0 \mid x_\tau]$.

We now turn to the formal statements and proofs of the technical lemmas, followed by the stage-by-stage analysis.

C. Technical Lemmas and Preliminary Derivations

C.1. Technical Lemmas and Their Proofs

Lemma 1 (Noise Projection Bound)**.** *For the spurious dense noise $\xi \sim \mathcal{N}(0, \sigma_\xi^2 \mathbf{I}_{d_1})$, where the variance satisfies $\omega\Big(\frac{1}{d_1}\Big) \leq \sigma_\xi^2 \leq O\big(\frac{1}{d}\big)$, the following holds with high probability $1 - e^{-\Omega(d_1)}$:*

$$|\langle w_i, \xi \rangle| \leq O\bigg(\frac{\|w_i\|_2}{d^{(1+c_0)/2}}\bigg), \quad \forall i \in [m]. \tag{13}$$

Proof. Since $\xi \sim \mathcal{N}(0, \sigma_\xi^2 \mathbf{I}_{d_1})$ is isotropic Gaussian, for any fixed direction $u \in \mathbb{R}^{d_1}$ the projection is a one-dimensional Gaussian:

$$\langle u, \xi \rangle \sim \mathcal{N}\big(0, \sigma_\xi^2 \|u\|_2^2\big)\,. \tag{14}$$

By standard Gaussian tail bounds, for all $j \in [d_1]$,

$$\Pr_\xi\bigg[|\langle \mathbf{M}_j, \xi \rangle| \leq O\bigg(\frac{1}{d^{(1+c_0)/2}}\bigg)\bigg] \geq 1 - e^{-\Omega(d_1)}, \tag{15}$$

and analogously for the orthogonal complement directions $\{\mathbf{M}_j^\perp\}_{j \in [d_1] \setminus [d]}$. Decompose w_i along the orthonormal basis induced by $\mathbf{M}$ and $\mathbf{M}^\perp$:

$$\langle w_i, \xi \rangle = \sum_{j \in [d]} \langle w_i, \mathbf{M}_j \rangle \langle \mathbf{M}_j, \xi \rangle + \sum_{j \in [d_1] \setminus [d]} \langle w_i, \mathbf{M}_j^\perp \rangle \langle \mathbf{M}_j^\perp, \xi \rangle. \tag{16}$$

By the Cauchy–Schwarz inequality, conditioning on the high-probability event above,

$$|\langle w_i, \xi \rangle| \leq \Big(\sum_{j \in [d]} |\langle w_i, \mathbf{M}_j \rangle|^2\Big)^{1/2} \cdot O\bigg(\frac{1}{d^{(1+c_0)/2}}\bigg) + \Big(\sum_{j \in [d_1] \setminus [d]} |\langle w_i, \mathbf{M}_j^\perp \rangle|^2\Big)^{1/2} \cdot O\bigg(\frac{1}{d^{(1+c_0)/2}}\bigg)\,. \tag{17}$$

Since $\|\mathbf{M}\mathbf{M}^\top w_i\|_2^2 + \|\mathbf{M}^\perp \mathbf{M}^{\perp\top} w_i\|_2^2 = \|w_i\|_2^2$, we obtain

$$|\langle w_i, \xi \rangle| \leq O\bigg(\frac{\|w_i\|_2}{d^{(1+c_0)/2}}\bigg)\,. \tag{18}$$

Thus, the lemma holds. □

Lemma 2 (Asymptotic Threshold for Indicator Activation)**.** *Let* $\sigma = \sqrt{\tau}\,\|w_i\|_2$*,* $\phi(u) = (2\pi)^{-1/2} e^{-u^2/2}$*, and assume that for some* $0 < c < 1$

$$\frac{\sigma^2}{\|w_i\|_2}\,\phi\!\left(\frac{b_i}{\sigma}\right) \;\geq\; d^{-c}. \tag{19}$$

Then the time parameter τ *satisfies*

$$\tau = \Theta\!\left(\frac{b_i^2}{\|w_i\|_2^2\,\log d}\right), \qquad d \to \infty. \tag{20}$$

Proof. Starting from

$$\frac{\sigma^2}{\|w_i\|_2}\,(2\pi)^{-1/2}\,\exp\!\left(-\frac{b_i^2}{2\sigma^2}\right) \;\geq\; d^{-c}, \tag{21}$$

multiplying both sides by $\sqrt{2\pi}\,\|w_i\|_2$ gives

$$\sigma^2 \exp\!\left(-\frac{b_i^2}{2\sigma^2}\right) \;\geq\; K, \qquad K = \sqrt{2\pi}\,\|w_i\|_2\,d^{-c}. \tag{22}$$

Let

$$a = \frac{b_i^2}{2}, \qquad y = \frac{a}{\sigma^2}, \qquad \sigma^2 = \frac{a}{y}. \tag{23}$$

The inequality becomes

$$\frac{a}{y}\,e^{-y} \geq K. \tag{24}$$

Equivalently,

$$y\,e^{y} \leq \frac{a}{K}. \tag{25}$$

Using the definition of the Lambert W function,

$$y \leq W\!\left(\frac{a}{K}\right), \qquad \sigma^2 \leq \frac{a}{W(a/K)}. \tag{26}$$

Thus

$$\tau \leq \frac{a}{\|w_i\|_2^2\,W(a/K)}. \tag{27}$$

Substitute

$$a = \frac{b_i^2}{2}, \qquad K = \sqrt{2\pi}\,\|w_i\|_2\,d^{-c}, \tag{28}$$

and define

$$x = \frac{b_i^2 d^c}{2\sqrt{2\pi}\,\|w_i\|_2}. \tag{29}$$

Then

$$\tau \leq \frac{b_i^2}{2\,\|w_i\|_2^2\,W(x)}. \tag{30}$$

For large d, $x = \Theta(d^c) \to \infty$. Using the expansion

$$W(x) = \log x - \log\log x + \frac{\log\log x}{\log x} + O\!\left(\frac{(\log\log x)^2}{(\log x)^2}\right), \tag{31}$$

and the identities

$$\log x = c\log d + O(1), \qquad \log\log x = \log\log d + O(1), \tag{32}$$

we obtain

$$W(x) = c\log d - \log\log d + O\!\left(\frac{\log\log d}{\log d}\right). \tag{33}$$

Therefore,

$$\frac{1}{W(x)} = \Theta\left(\frac{1}{\log d}\right). \tag{34}$$

Substituting this into the upper bound for τ yields

$$\tau = \Theta\left(\frac{b_i^2}{\|w_i\|_2^2 \log d}\right) \tag{35}$$

□

Lemma 3 (Conditional Lower Tail for a Gaussian Component Given a Large Sum)**.** *Let $X_1, \ldots, X_k$ be jointly Gaussian with $X_i \sim \mathcal{N}(0, \tau)$, and let*

$$Y = \sum_{i=1}^{k} X_i, \tag{36}$$

so that $Y \sim \mathcal{N}(0, k\tau)$. For any threshold $b_i > 0$ and any constant $0 < c < 1$, the conditional distribution of X_i under the event $Y > b_i$ satisfies

$$\Pr\left(X_i > c\,\frac{b_i}{k} \,\middle|\, Y > b_i\right) \geq 1 - \exp\left(-\frac{(1-c)^2}{2}\,\frac{(b_i/k)^2}{\tau(1-\frac{1}{k})}\right). \tag{37}$$

Proof. Since $X_1, \ldots, X_k$ are jointly Gaussian with $X_i \sim \mathcal{N}(0, \tau)$ and

$$Y = \sum_{i=1}^{k} X_i, \tag{38}$$

the pair (X_i, Y) is bivariate Gaussian. We have

$$\mathrm{Cov}(X_i, Y) = \mathrm{Cov}\Big(X_i, \sum_{j=1}^{k} X_j\Big) = \sum_{j=1}^{k} \mathrm{Cov}(X_i, X_j) = \mathrm{Var}(X_i) = \tau, \tag{39}$$

and

$$\mathrm{Var}(Y) = \sum_{j=1}^{k} \mathrm{Var}(X_j) = k\tau. \tag{40}$$

Therefore the conditional law of X_i given $Y = y$ is Gaussian with

$$X_i \mid Y = y \sim \mathcal{N}\left(\frac{\mathrm{Cov}(X_i, Y)}{\mathrm{Var}(Y)}\, y,\ \mathrm{Var}(X_i) - \frac{\mathrm{Cov}(X_i, Y)^2}{\mathrm{Var}(Y)}\right) = \mathcal{N}\left(\frac{y}{k},\ \tau\Big(1 - \frac{1}{k}\Big)\right). \tag{41}$$

Now condition on the event $Y > b_i$. By the tower property,

$$\mathbb{E}[X_i \mid Y > b_i] = \mathbb{E}\big[\mathbb{E}[X_i \mid Y] \,\big|\, Y > b_i\big] = \mathbb{E}\big[Y/k \,\big|\, Y > b_i\big] = \frac{1}{k}\,\mathbb{E}[Y \mid Y > b_i]. \tag{42}$$

Since $Y \sim \mathcal{N}(0, k\tau)$, the conditional distribution of Y given $Y > b_i$ is a one-dimensional Gaussian truncated to (b_i, ∞), and

$$\mathbb{E}[Y \mid Y > b_i] \geq b_i. \tag{43}$$

Hence

$$\mathbb{E}[X_i \mid Y > b_i] = \frac{1}{k}\,\mathbb{E}[Y \mid Y > b_i] \geq \frac{b_i}{k}. \tag{44}$$

The conditional variance does not depend on Y, so

$$\mathrm{Var}(X_i \mid Y = y) = \tau\Big(1 - \frac{1}{k}\Big) \tag{45}$$

for every y, and therefore

$$\mathrm{Var}(X_i \mid Y > b_i) = \tau\Big(1 - \frac{1}{k}\Big). \tag{46}$$

Thus $X_i \mid Y > b_i$ is a mixture of Gaussians of the form $\mathcal{N}(y/k,\, \tau(1-1/k))$ with $y > b_i$, so its mean is at least b_i/k and its variance is $\tau(1-1/k)$.

Fix a constant c with $0 < c < 1$. Write

$$\mu_i = \mathbb{E}[X_i \mid Y > b_i], \qquad \sigma_i^2 = \mathrm{Var}(X_i \mid Y > b_i) = \tau\Big(1 - \frac{1}{k}\Big). \tag{47}$$

We have $\mu_i \geq b_i/k$. Consider any Gaussian random variable $Z \sim \mathcal{N}(\mu, \sigma^2)$ and any threshold $u < \mu$. The Gaussian tail bound

$$\Phi(-x) \leq \exp\left(-\frac{x^2}{2}\right) \quad \text{for all } x > 0 \tag{48}$$

implies

$$\Pr(Z \leq u) = \Pr\left(\frac{Z-\mu}{\sigma} \leq \frac{u-\mu}{\sigma}\right) = \Phi\left(\frac{u-\mu}{\sigma}\right) \leq \exp\left(-\frac{(\mu-u)^2}{2\sigma^2}\right), \tag{49}$$

and therefore

$$\Pr(Z > u) = 1 - \Pr(Z \leq u) \geq 1 - \exp\left(-\frac{(\mu-u)^2}{2\sigma^2}\right). \tag{50}$$

Apply this with

$$\mu = \mu_i, \qquad \sigma^2 = \sigma_i^2 = \tau\Big(1 - \frac{1}{k}\Big), \qquad u = c\,\frac{b_i}{k}, \tag{51}$$

and use $\mu_i \geq b_i/k$. Then $\mu_i - u \geq (1-c)\, b_i/k$, so

$$\Pr\left(X_i > c\,\frac{b_i}{k} \,\middle|\, Y > b_i\right) \geq 1 - \exp\left(-\frac{(1-c)^2}{2}\,\frac{(b_i/k)^2}{\tau(1-\frac{1}{k})}\right). \tag{52}$$

This is exactly the desired inequality. □

Lemma 4 (Two–regime Gaussian threshold behavior)**.** *Let $X_1 \sim \mathcal{N}(a_1, \tau)$ and $X_2 \sim \mathcal{N}(a_2, \tau)$ be two Gaussian random variables with the same variance $\tau > 0$. Assume the decision threshold b_ℓ lies between the means:*

$$a_1 < b_\ell < a_2. \tag{53}$$

Define the signed gaps (distances to the threshold) as:

$$\Delta_1 = b_\ell - a_1 > 0, \qquad \Delta_2 = b_\ell - a_2 < 0. \tag{54}$$

Then, for any $\delta \in (0, 1/2)$, the following two regimes hold:

(A) Separable Regime (Small τ): If the noise level satisfies

$$\sqrt{\tau} \leq \min\left\{\frac{\Delta_1}{\sqrt{2\log(1/\delta)}}, \frac{|\Delta_2|}{\sqrt{2\log(1/\delta)}}\right\}, \tag{55}$$

then the error probabilities are bounded by δ:

$$\Pr(X_1 \geq b_\ell) \leq \delta, \qquad \Pr(X_2 \geq b_\ell) \geq 1 - \delta. \tag{56}$$

(B) Mixed Regime (Large τ): If the noise level satisfies

$$\sqrt{\tau} \geq \max\left\{\frac{|\Delta_1|}{\sqrt{2\pi}\,\delta}, \frac{|\Delta_2|}{\sqrt{2\pi}\,\delta}\right\}, \tag{57}$$

then the probabilities are close to random guessing:

$$\left|\Pr(X_1 \geq b_\ell) - \tfrac{1}{2}\right| \leq \delta, \qquad \left|\Pr(X_2 \geq b_\ell) - \tfrac{1}{2}\right| \leq \delta. \tag{58}$$

Proof. For $X_j \sim \mathcal{N}(a_j, \tau)$, we express the tail probability in terms of the standard normal CDF $\Phi(\cdot)$:

$$\Pr(X_j \geq b_\ell) = 1 - \Phi\left(\frac{b_\ell - a_j}{\sqrt{\tau}}\right) = 1 - \Phi\left(\frac{\Delta_j}{\sqrt{\tau}}\right). \tag{59}$$

Proof of (A). Define the normalized distances $\zeta_1 = \frac{\Delta_1}{\sqrt{\tau}}$ and $\zeta_2 = \frac{-\Delta_2}{\sqrt{\tau}}$. From the condition in Regime (A), we have:

$$\zeta_1 \geq \sqrt{2\log(1/\delta)} \quad \text{and} \quad \zeta_2 \geq \sqrt{2\log(1/\delta)}. \tag{60}$$

We utilize the standard Gaussian tail bound (Chernoff bound), which states that for any $\zeta > 0$, $1 - \Phi(\zeta) \leq \exp(-\zeta^2/2)$. Applying this to X_1:

$$\Pr(X_1 \geq b_\ell) = 1 - \Phi(\zeta_1) \leq \exp\left(-\frac{\zeta_1^2}{2}\right) \leq \exp\left(-\frac{2\log(1/\delta)}{2}\right) = \delta. \tag{61}$$

Similarly for X_2, noting that $\Delta_2 = -\sqrt{\tau}\zeta_2$:

$$\Pr(X_2 \geq b_\ell) = 1 - \Phi(-\zeta_2) = \Phi(\zeta_2) \geq 1 - \exp\left(-\frac{\zeta_2^2}{2}\right) \geq 1 - \delta. \tag{62}$$

Proof of (B). Define the standardized gaps $\xi_j = \frac{\Delta_j}{\sqrt{\tau}}$. The condition in Regime (B) implies:

$$|\xi_j| \leq \sqrt{2\pi}\,\delta, \quad \text{for } j \in \{1, 2\}. \tag{63}$$

Recall that the standard normal PDF $\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$ is bounded by $\frac{1}{\sqrt{2\pi}}$. By the Mean Value Theorem, for any ξ:

$$\left|\Phi(\xi) - \tfrac{1}{2}\right| = \left|\Phi(\xi) - \Phi(0)\right| \leq \sup_{u \in \mathbb{R}} |\phi(u)| \cdot |\xi| = \frac{|\xi|}{\sqrt{2\pi}}. \tag{64}$$

From (59), we have $\Pr(X_j \geq b_\ell) = \Phi(-\xi_j)$. Thus:

$$\left|\Pr(X_j \geq b_\ell) - \tfrac{1}{2}\right| = \left|\Phi(-\xi_j) - \Phi(0)\right| \leq \frac{|-\xi_j|}{\sqrt{2\pi}} \leq \frac{\sqrt{2\pi}\,\delta}{\sqrt{2\pi}} = \delta. \tag{65}$$

This completes the proof. □

Lemma 5 (Gaussian logistic moment (standard approximation))**.** *Let $X \sim \mathcal{N}(\mu, s^2)$ and $\sigma(x) = \frac{1}{1+e^{-x}}$. Then*

$$\mathbb{E}[\sigma(X)] = \int_{\mathbb{R}} \sigma(x)\, \frac{1}{\sqrt{2\pi}s} \exp\left(-\frac{(x-\mu)^2}{2s^2}\right) dx, \tag{66}$$

and admits the widely used approximation

$$\mathbb{E}[\sigma(X)] \approx \sigma\left(\frac{\mu}{\sqrt{1 + \frac{\pi^2}{8}s^2}}\right). \tag{67}$$

Proof. Start from the definition

$$\mathbb{E}[\sigma(X)] = \int_{\mathbb{R}} \frac{1}{1+e^{-x}}\, \frac{1}{\sqrt{2\pi}s} \exp\left(-\frac{(x-\mu)^2}{2s^2}\right) dx. \tag{68}$$

There is no elementary closed form for this integral. To obtain an analytic expression, use the standard approximation that matches the logistic link to a probit link:

$$\sigma(x) \approx \Phi(\kappa x), \qquad \kappa = \sqrt{\frac{8}{\pi^2}}. \tag{69}$$

Under this substitution,

$$\mathbb{E}[\sigma(X)] \approx \mathbb{E}[\Phi(\kappa X)]. \tag{70}$$

Let $Z \sim \mathcal{N}(0,1)$ independent of X. Using $\Phi(t) = \Pr(Z \le t)$,

$$\mathbb{E}[\Phi(\kappa X)] = \mathbb{E}\big[\Pr(Z \le \kappa X \mid X)\big] = \Pr(Z \le \kappa X) = \Pr(\kappa X - Z \ge 0). \tag{71}$$

Since $\kappa X - Z$ is Gaussian with mean $\kappa\mu$ and variance $\kappa^2 s^2 + 1$, we have

$$\Pr(\kappa X - Z \ge 0) = \Phi\left(\frac{\kappa\mu}{\sqrt{1+\kappa^2 s^2}}\right). \tag{72}$$

Applying the inverse substitution $\Phi(\kappa u) \approx \sigma(u)$ gives

$$\mathbb{E}[\sigma(X)] \approx \sigma\left(\frac{\mu}{\sqrt{1+\kappa^{-2} s^2}}\right) = \sigma\left(\frac{\mu}{\sqrt{1+\frac{\pi^2}{8} s^2}}\right), \tag{73}$$

which completes the proof. □

Definition 3 (Neuron Characterization)**.** Let us define a few notations to characterize each neuron $w_i^{(t)}$'s behavior. For every constant $c_0 \in (0,1)$ and $\gamma \in (0, 0.1)$, by choosing $c_1 = 2 + 2(1-\gamma)c_0$ and $c_2 = \gamma c_0$, we define:

1. Let $\mathcal{S}_{j,\text{sure}}^{(t)} \subseteq [m]$ be those neurons $i \in [m]$ satisfying

- $(\frac{1}{n}\sum_{i=1}^n \langle w_i^{(t)}, \mathbf{M}_j\rangle)^2 \ge \frac{(c_1+c_2)\log d}{d}\|\mathbf{M}\mathbf{M}^\top w_i^{(t)}\|_2^2$
- $(\frac{1}{n}\sum_{i=1}^n \langle w_i^{(t)}, \mathbf{M}_{j'}\rangle)^2 < \frac{(c_1-c_2)\log d}{d}\|\mathbf{M}\mathbf{M}^\top w_i^{(t)}\|_2^2$

2. Let $\mathcal{S}_{j,\text{pot}}^{(t)} \subseteq [m]$ be those neurons $i \in [m]$ satisfying

- $\langle w_i^{(t)}, \mathbf{M}_j\rangle^2 \ge \frac{(c_1-c_2)\log d}{d}\|\mathbf{M}\mathbf{M}^\top w_i^{(t)}\|_2^2$

Lemma 6 (Geometry at initialization)**.** *We initialize the parameters by $w_i^{(0)} \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I}_{d_1})$, where $\sigma_0^2 = \Theta\left(\frac{1}{d_1 poly(d)}\right)$. We have with probability $\ge 1 - o(1/d^3)$ over the random initialization, for all $j \in [d]$:*

$$\left|\mathcal{S}_{j,sure}^{(0)}\right| = \Omega\left(d^{\frac{\gamma}{4}c_0}\right) =: \Xi_1$$
$$\left|\mathcal{S}_{j,pot}^{(0)}\right| \le O\left(d^{2\gamma c_0}\right) =: \Xi_2$$

Proof. If g is standard Gaussian, then for every $t > 0$,

$$\frac{1}{\sqrt{2\pi}}\frac{(t)}{t^2+1}e^{-t^2/2} < \Pr_{g\sim\mathcal{N}(0,1)}[g > t] < \frac{1}{\sqrt{2\pi}}\frac{1}{(t)}e^{-t^2/2}. \tag{74}$$

We initialize the parameters by $w_i^{(0)} \sim \mathcal{N}(0, \sigma_0^2\mathbf{I}_{d_1})$, where $\sigma_0^2 = \Theta\left(\frac{1}{d_1 \text{poly}(d)}\right)$. We have $\frac{1}{n}\sum_{i=1}^n \langle w_i^{(0)}, \mathbf{M}_i\rangle \sim \mathcal{N}\left(0, \frac{\sigma_0^2}{n}\right)$.

Therefore, for every $i \in m$ and $j \in d$, we have

$$\begin{aligned} p_1 &= \Pr\left[\left(\frac{1}{n}\sum_{i=1}^n \langle w_i^{(0)}, \mathbf{M}_j\rangle\right)^2 \ge (c_1+c_2)\frac{\sigma_0^2}{n}\log d\right] \\ &= \Theta\left(\frac{1}{\log d}\right)\cdot\frac{1}{d^{(c_1+c_2)/2}} \\ &= \Theta\left(\frac{1}{\sqrt{\log d}}\right)\cdot\frac{1}{d\cdot d^{(1-\gamma/2)c_0}} \end{aligned} \tag{75}$$

$$
\begin{aligned}
p_2 &= \Pr\left[\left(\frac{1}{n}\sum_{i=1}^{n}\langle w_i^{(0)}, \mathbf{M}_{j'}\rangle\right)^2 \geq (c_1-c_2)\frac{\sigma_0^2}{n}\log d\right] \\
&= \Theta\left(\frac{1}{\log d}\right)\cdot\frac{1}{d^{(c_1-c_2)/2}} \\
&= \Theta\left(\frac{1}{\sqrt{\log d}}\right)\cdot\frac{1}{d\cdot d^{(1-3\gamma/2)c_0}}
\end{aligned} \tag{76}
$$

Let $\mathcal{S}_{j,\text{sure}}^{(0)} \subseteq [m]$ be those neurons $i \in [m]$ satisfying

- $(\frac{1}{n}\sum_{i=1}^{n}\langle w_i^{(0)}, \mathbf{M}_j\rangle)^2 \geq \frac{(c_1+c_2)\log d}{d}\|\mathbf{M}\mathbf{M}^\top w_i^{(0)}\|_2^2$
- $(\frac{1}{n}\sum_{i=1}^{n}\langle w_i^{(0)}, \mathbf{M}_{j'}\rangle)^2 < \frac{(c_1-c_2)\log d}{d}\|\mathbf{M}\mathbf{M}^\top w_i^{(0)}\|_2^2$

By concentration with respect to all m choices of $i \in [m]$, we know with probability at least $1 - o\left(\frac{1}{d^3}\right)$ it satisfies $\left|\mathcal{S}_{j,\text{sure}}^{(0)}\right| = \Omega\left(d^{\frac{\gamma}{4}c_0}\right)$.

Let $\mathcal{S}_{j,\text{pot}}^{(0)} \subseteq [m]$ be those neurons $i \in [m]$ satisfying

- $\langle w_i^{(0)}, \mathbf{M}_j\rangle^2 \geq \frac{(c_1-c_2)\log d}{d}\|\mathbf{M}\mathbf{M}^\top w_i^{(0)}\|_2^2$

By concentration with respect to all m choices of $i \in [m]$, we know with probability at least $1 - o\left(\frac{1}{d^3}\right)$ it satisfies $\left|\mathcal{S}_{j,\text{pot}}^{(0)}\right| = O\left(d^{2\gamma c_0}\right)$.

More details of the proof can be found in Lemma B.2 of (Allen-Zhu & Li, 2022). □

Lemma 7. *With high probability* $1 - \frac{1}{\text{poly}(d)}$*, for every* $i \in [m]$*, the following holds:*

$$
\Pr\left[(\frac{1}{2n}\sum_{i=1}^{n}\langle w_i^{(0)}, \mathbf{M}_j\rangle - \langle w_i^{(0)}, \mathbf{M}_{j'}\rangle)^2 \geq \frac{1}{d}\frac{\sigma_0^2}{2n}\log d\right] \geq 1 - O(\frac{1}{\sqrt{d}}) \tag{77}
$$

Lemma 8 (Concentration bound for empirical loss and gradients)**.** *There exist* $N \geq \text{poly}(d)$ *for some sufficiently large polynomial and all* $\|w_i\|_2 \leq O(d)$*,* $i \in [m]$*, it satisfies*

$$
\left|\frac{1}{N}\sum_{p\in[N]}\tfrac{1}{2}\|g(x_{\tau,p}) - x_{0,p}\|_2^2 - \mathbb{E}_{x_0,\tau}[\tfrac{1}{2}\|g(x_\tau) - x_0\|_2^2]\right| \leq O\left(\frac{1}{d}\right) \tag{78}
$$

$$
\left\|\frac{1}{N}\sum_{p\in[N]}\nabla_{w_i}L_{\text{DM}}(x_{0,p}, x_{\tau,p}) - \mathbb{E}_{x_0,\tau}[\nabla_{w_i}L_{\text{DM}}]\right\|_2 \leq O\left(\frac{1}{d}\right) \tag{79}
$$

Proof. The proof can be done by trivial VC dimension or Rademacher complexity arguments similarly to Lemma A.2 in (Allen-Zhu & Li, 2022). □

C.2. Preliminary Analytical Derivations

This section establishes the notation and derives the update dynamics that underlie our theoretical analysis.

C.2.1. NETWORK ARCHITECTURE

We consider a one-hidden-layer ReLU network $g : \mathbb{R}^{d_1} \to \mathbb{R}^{d_1}$ as the denoiser:

$$g(x) = V^\top \sigma(Wx), \tag{80}$$

where $W = [w_1, \dots, w_m]^\top \in \mathbb{R}^{m \times d_1}$ and $V = [v_1, \dots, v_m]^\top \in \mathbb{R}^{m \times d_1}$ are the input and output weight matrices, m is the number of hidden neurons, and $\sigma(\cdot) = \max(\cdot, 0)$ denotes the ReLU activation applied element-wise.

The weights are initialized as $w_i^{(0)}, v_i^{(0)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_0^2 I_{d_1})$ for $i \in [m]$.

The clean signal follows the multi-scale sparse coding model:

$$x_0 = \alpha_1 \mathbf{M}_1 z_1 + \alpha_2 \mathbf{M}_2 z_2, \tag{81}$$

where $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{d_1 \times d}$ are orthonormal dictionaries with $\mathbf{M}_1^\top \mathbf{M}_2 = 0$, and $z_1, z_2 \in \{-1, 0, 1\}^d$ are sparse codes. The scale separation $\alpha_1 = \Theta(1) \gg \alpha_2 = \Theta(1/d^{c_0})$ captures the hierarchy between coarse and fine-grained features.

During training, the noisy observation is

$$x_\tau = x_0 + \tau\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I_{d_1}), \tag{82}$$

where $\tau \sim \text{Unif}[\tau_{\min}, \tau_{\max}]$ is the noise level.

C.2.2. GRADIENT COMPUTATION

We train the network using stochastic gradient descent on the diffusion objective

$$L_{\text{DM}} = \mathbb{E}_{x_0, \tau, \epsilon} \tfrac{1}{2} \|g(x_\tau) - x_0\|_2^2. \tag{83}$$

The gradients with respect to the weight vectors are

$$\frac{\partial L_{\text{DM}}}{\partial w_i} = \langle v_i, g(x_\tau) - x_0 \rangle \, x_\tau \, \mathbf{1}\{w_i^\top x_\tau \geq 0\}, \tag{84}$$

$$\frac{\partial L_{\text{DM}}}{\partial v_i} = (w_i^\top x_\tau) \, (g(x_\tau) - x_0) \, \mathbf{1}\{w_i^\top x_\tau \geq 0\}, \tag{85}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function arising from the ReLU activation.

The parameter updates follow

$$w_i^{(t+1)} = w_i^{(t)} + \eta \, \langle v_i^{(t)}, x_0 - g^{(t)}(x_\tau) \rangle x_\tau \, \mathbf{1}\{w_i^{(t)\top} x_\tau \geq 0\}, \tag{86}$$

$$v_i^{(t+1)} = v_i^{(t)} + \eta \, (x_0 - g^{(t)}(x_\tau))(w_i^{(t)\top} x_\tau) \, \mathbf{1}\{w_i^{(t)\top} x_\tau \geq 0\}, \tag{87}$$

where $\eta > 0$ is the learning rate.

C.2.3. ALIGNMENT UPDATES

To characterize feature learning, we analyze the alignment between network weights and dictionary directions. For neuron i and dictionary column $\mathbf{M}_{kj}$ ($k \in \{1, 2\}$, $j \in [d]$), the alignment is defined as $\langle w_i^{(t)}, \mathbf{M}_{kj} \rangle$.

Projecting the update rule (86) onto $\mathbf{M}_{1j}$ yields

$$\langle w_i^{(t+1)}, \mathbf{M}_{1j} \rangle = \langle w_i^{(t)}, \mathbf{M}_{1j} \rangle + \eta \, \langle v_i^{(t)}, x_0 - g^{(t)}(x_\tau) \rangle \langle x_\tau, \mathbf{M}_{1j} \rangle \, \mathbf{1}\{w_i^{(t)\top} x_\tau \geq 0\}. \tag{88}$$

Using the data model (81) and the orthonormality $\mathbf{M}_1^\top \mathbf{M}_1 = I_d$, $\mathbf{M}_1^\top \mathbf{M}_2 = 0$, we have

$$\langle x_\tau, \mathbf{M}_{1j} \rangle = \alpha_1 z_1^j + \tau \langle \epsilon, \mathbf{M}_{1j} \rangle, \tag{89}$$

where $\langle \epsilon, \mathbf{M}_{1j} \rangle \sim \mathcal{N}(0,1)$.

When $g^{(t)}(x_\tau)$ has not yet learned the feature $\mathbf{M}_{1j}$, the residual $x_0 - g^{(t)}(x_\tau)$ contains a significant $\mathbf{M}_{1j}$ component. Substituting and taking expectation over $z_1^j \in \{-1, 0, 1\}$, the coupled dynamics become

$$\langle w_i^{(t+1)}, \mathbf{M}_{1j} \rangle \approx \langle w_i^{(t)}, \mathbf{M}_{1j} \rangle + \eta\, \alpha_1^2\, \langle v_i^{(t)}, \mathbf{M}_{1j} \rangle\, \mathrm{Pr}_{1j} \pm \mathrm{Err}_t, \tag{90}$$

$$\langle v_i^{(t+1)}, \mathbf{M}_{1j} \rangle \approx \langle v_i^{(t)}, \mathbf{M}_{1j} \rangle + \eta\, \alpha_1^2\, \langle w_i^{(t)}, \mathbf{M}_{1j} \rangle\, \mathrm{Pr}_{1j} \pm \mathrm{Err}_t, \tag{91}$$

where $\mathrm{Pr}_{1j} = \Pr(z_1^j \neq 0) \cdot \Pr(\mathbf{1}\{w_i^{(t)\top} x_\tau \geq 0\} = 1)$, and Err_t collects contributions from Gaussian noise, the weak $\mathbf{M}_2$ component, and cross-terms.

These dynamics admit the matrix form

$$\begin{bmatrix} \langle w_i^{(t+1)}, \mathbf{M}_{1j} \rangle \\ \langle v_i^{(t+1)}, \mathbf{M}_{1j} \rangle \end{bmatrix} = \begin{bmatrix} 1 & a_t \\ a_t & 1 \end{bmatrix} \begin{bmatrix} \langle w_i^{(t)}, \mathbf{M}_{1j} \rangle \\ \langle v_i^{(t)}, \mathbf{M}_{1j} \rangle \end{bmatrix} \pm \mathrm{Err}_t, \tag{92}$$

where $a_t = \Theta(\eta \alpha_1^2 / d)$ when the indicator is active. The matrix has eigenvalues $1 + a_t$ and $1 - a_t$, so the dominant eigenmode grows as $(1 + a_t)^t$. For neurons with favorable initialization satisfying $\langle w_i^{(0)}, \mathbf{M}_{1j} \rangle + \langle v_i^{(0)}, \mathbf{M}_{1j} \rangle \geq \Omega(\sigma_0 / \sqrt{d})$, the alignment grows exponentially until reaching $\Theta(1)$.

Analogous dynamics hold for $\mathbf{M}_{2j}$, but with growth rate $\tilde{a}_t = \Theta(\eta \alpha_2^2 / d)$, slower by a factor of $\alpha_2^2 / \alpha_1^2 = \Theta(1/d^{2c_0})$. This scale separation underlies the two-stage learning: neurons first align with $\mathbf{M}_1$ before developing $\mathbf{M}_2$ alignment.

D. Stage 0: Feature Initialization

In this section, we analyze Stage 0 of the diffusion model training dynamics, corresponding to the early training iterations $t \leq T_0$, where $T_0 = \Theta\Big(\frac{d \log d}{\eta}\Big)$ denotes the time scale at which the energy of each neuron has not yet significantly deviated from its random initialization, i.e., $\frac{\|w_i^{(T_0)}\|_2^2 + \|v_i^{(T_0)}\|_2^2}{2} \geq \|w_i^{(0)}\|_2^2 + \|v_i^{(0)}\|_2^2$ for all $i \in [m]$. Throughout this stage, we set $b_i^{(t)} = 0$, so that the training dynamics are governed purely by the denoising gradients induced by the corrupted observations.

For every neuron $i \in [m]$, the weights w_i and v_i exhibit a pronounced growth in alignment with $\mathbf{M}_1$, while the growth of alignment with $\mathbf{M}_2$ remains strictly smaller. Moreover, the alignment along directions orthogonal to the signal subspace exhibits only negligible increase throughout this stage.

Since $x_0 = \alpha_1 \mathbf{M}_1 z_1 + \alpha_2 \mathbf{M}_2 z_2$ with $\alpha_1 = \Theta(1)$ and $\alpha_2 = \Theta(1/d^{c_0})$, the gradient contributions from $\mathbf{M}_1$ dominate those from $\mathbf{M}_2$. Specifically, the expected gradient along $\mathbf{M}_{1j}$ scales as $\mathbb{E}[z_{1j}^2] = \Theta(1/d)$, while the effective contribution from $\mathbf{M}_{2j}$ is suppressed by the factor $\alpha_2^2 = \Theta(1/d^{2c_0})$.

We now project the SGD updates onto the feature directions $\mathbf{M}_{1j}$ and $\mathbf{M}_{2j}$. The projected updates become:

$$\langle w_i^{(t+1)}, \mathbf{M}_{kj} \rangle = \langle w_i^{(t)}, \mathbf{M}_{kj} \rangle + \eta\, \langle v_i^{(t)}, \mathbf{M}_{kj} \rangle\, \alpha_k z_{kj}\, \mathbf{1}\{w_i^{(t)\top} x_\tau \geq b_i^{(t)}\} \pm \mathrm{Err}_t, \tag{93}$$

$$\langle v_i^{(t+1)}, \mathbf{M}_{kj} \rangle = \langle v_i^{(t)}, \mathbf{M}_{kj} \rangle + \eta\, \langle w_i^{(t)}, \mathbf{M}_{kj} \rangle\, \alpha_k z_{kj}\, \mathbf{1}\{w_i^{(t)\top} x_\tau \geq b_i^{(t)}\} \pm \mathrm{Err}_t, \tag{94}$$

for $k \in \{1, 2\}$.

With $b_i^{(t)} = 0$ in Stage 0, taking expectation over $z_{kj} \sim \mathrm{Bernoulli}(1/d)$ and the gating event yields:

$$\mathbb{E}\left[\alpha_k z_{kj} \cdot \mathbf{1}\{w_i^{(t)\top} x_\tau \geq 0\}\right] = \Theta\left(\frac{\alpha_k}{d}\right). \tag{95}$$

The coupled dynamics can be written as:

$$\begin{bmatrix} \langle w_i^{(t+1)}, \mathbf{M}_{kj} \rangle \\ \langle v_i^{(t+1)}, \mathbf{M}_{kj} \rangle \end{bmatrix} = \begin{bmatrix} 1 & a_k \\ a_k & 1 \end{bmatrix} \begin{bmatrix} \langle w_i^{(t)}, \mathbf{M}_{kj} \rangle \\ \langle v_i^{(t)}, \mathbf{M}_{kj} \rangle \end{bmatrix} \pm \mathrm{Err}_t, \tag{96}$$

where

$$a_1 = \Theta\left(\frac{\eta\alpha_1}{d}\right) = \Theta\left(\frac{\eta}{d}\right), \qquad a_2 = \Theta\left(\frac{\eta\alpha_2}{d}\right) = \Theta\left(\frac{\eta}{d^{1+c_0}}\right). \tag{97}$$

Iterating the recursion yields:

$$\langle w_i^{(t)}, \mathbf{M}_{kj}\rangle = (1+a_k)^t \ \frac{\langle w_i^{(0)}, \mathbf{M}_{kj}\rangle + \langle v_i^{(0)}, \mathbf{M}_{kj}\rangle}{2}. \tag{98}$$

By symmetry of the coupled dynamics, the same closed-form holds for $\langle v_i^{(t)}, \mathbf{M}_{kj}\rangle$.

D.1. Lower Bound of M_1 Alignment

We first establish a lower bound for $\|\mathbf{M}_1\mathbf{M}_1^\top w_i^{(t)}\|_2^2$. From Eq. (98), we have:

$$\|\mathbf{M}_1\mathbf{M}_1^\top w_i^{(t)}\|_2^2 = (1+a_1)^{2t} \cdot \frac{\|\mathbf{M}_1\mathbf{M}_1^\top w_i^{(0)}\|_2^2 + \|\mathbf{M}_1\mathbf{M}_1^\top v_i^{(0)}\|_2^2}{4}. \tag{99}$$

By symmetry, the same bound holds for $\|\mathbf{M}_1\mathbf{M}_1^\top v_i^{(t)}\|_2^2$.

For directions orthogonal to $\mathbf{M}_1$, we analyze the projection of $w_i^{(t)}$ onto the null space of $\mathbf{M}_1$. Since the clean signal $x_0 = \alpha_1\mathbf{M}_1 z_1 + \alpha_2\mathbf{M}_2 z_2$ lies entirely in the span of $\mathbf{M}_1$ and $\mathbf{M}_2$, the gradient contributions along directions orthogonal to both $\mathbf{M}_1$ and $\mathbf{M}_2$ arise solely from the noise term $\tau\epsilon$.

Projecting the update rule onto the orthogonal complement, we obtain:

$$\langle w_i^{(t+1)}, u\rangle = \langle w_i^{(t)}, u\rangle + \eta\,\langle v_i^{(t)}, x_0 - g^{(t)}(x_\tau)\rangle\langle x_\tau, u\rangle\,\mathbf{1}\{w_i^{(t)\top} x_\tau \geq 0\} \tag{100}$$

for any $u \perp \mathrm{span}(\mathbf{M}_1, \mathbf{M}_2)$. Since $\langle x_0, u\rangle = 0$ and $\langle x_\tau, u\rangle = \tau\langle\epsilon, u\rangle$, the update is driven purely by noise. Taking expectation over ϵ and using $\mathbb{E}[\epsilon] = 0$, the expected drift along orthogonal directions vanishes.

By concentration of Gaussian quadratic forms (Lemma 1), with high probability $1 - e^{-\Omega(d)}$,

$$\|(I - \mathbf{M}_1\mathbf{M}_1^\top)w_i^{(t)}\|_2^2 \leq \left(1 + \frac{1}{\mathrm{poly}(d)}\right)\|(I - \mathbf{M}_1\mathbf{M}_1^\top)w_i^{(0)}\|_2^2 \leq O\left(\frac{1}{d_1}\right)\|w_i^{(0)}\|_2^2 \tag{101}$$

for all $t \leq T_0$. This demonstrates that the components of $w_i^{(t)}$ orthogonal to $\mathbf{M}_1$ experience negligible growth during Stage 0, remaining at their initialization scale.

D.2. Lower Bound of M_2 Alignment

We next establish a lower bound for $\|\mathbf{M}_2\mathbf{M}_2^\top w_i^{(t)}\|_2^2$. From Eq. (98), we have:

$$\|\mathbf{M}_2\mathbf{M}_2^\top w_i^{(t)}\|_2^2 = (1+a_2)^{2t} \cdot \frac{\|\mathbf{M}_2\mathbf{M}_2^\top w_i^{(0)}\|_2^2 + \|\mathbf{M}_2\mathbf{M}_2^\top v_i^{(0)}\|_2^2}{4}. \tag{102}$$

By symmetry, the same bound holds for $\|\mathbf{M}_2\mathbf{M}_2^\top v_i^{(t)}\|_2^2$.

Similarly, for directions orthogonal to $\mathbf{M}_2$, the same analysis applies. Since $\langle x_0, u\rangle = 0$ for $u \perp \mathrm{span}(\mathbf{M}_1, \mathbf{M}_2)$, the gradient contributions along such directions arise solely from the noise term. With high probability $1 - e^{-\Omega(d)}$,

$$\|(I - \mathbf{M}_2\mathbf{M}_2^\top)w_i^{(t)}\|_2^2 \leq \left(1 + \frac{1}{\mathrm{poly}(d)}\right)\|(I - \mathbf{M}_2\mathbf{M}_2^\top)w_i^{(0)}\|_2^2 \leq O\left(\frac{1}{d_1}\right)\|w_i^{(0)}\|_2^2 \tag{103}$$

for all $t \leq T_0$.

This demonstrates that $\mathbf{M}_2$ alignment grows exponentially (albeit at a slower rate than $\mathbf{M}_1$ due to the suppression factor $a_2 = \Theta(\eta/d^{1+c_0})$), while orthogonal components remain essentially unchanged.

D.3. Alignment Growth for $i \in S_{1j,\text{sure}}$

For neurons $i \in S_{1j,\text{sure}}$, we prove the lower bound of $|\langle w_i^{(T_0)}, \mathbf{M}_{1j}\rangle|^2$:

$$\begin{aligned}
|\langle w_i^{(T_0)}, \mathbf{M}_{1j}\rangle|^2 &= (1+a_1)^{2T_0}\left(\frac{\langle w_i^{(0)}, \mathbf{M}_{1j}\rangle + \langle v_i^{(0)}, \mathbf{M}_{1j}\rangle}{2}\right)^2 \\
&\overset{(a)}{\geq} (1+a_1)^{2T_0}\cdot\frac{(c_1+c_2)\log d}{d}\cdot\frac{\|\mathbf{M}_1\mathbf{M}_1^\top w_i^{(0)}\|_2^2 + \|\mathbf{M}_1\mathbf{M}_1^\top v_i^{(0)}\|_2^2}{4} \\
&\overset{(b)}{=} \frac{(c_1+c_2)\log d}{d}\cdot\frac{\|\mathbf{M}_1\mathbf{M}_1^\top w_i^{(T_0)}\|_2^2 + \|\mathbf{M}_1\mathbf{M}_1^\top v_i^{(T_0)}\|_2^2}{2} \\
&\overset{(c)}{\geq} \frac{(c_1+c_2)\log d}{d}\cdot\frac{\|w_i^{(T_0)}\|_2^2 + \|v_i^{(T_0)}\|_2^2 - \|w_i^{(0)}\|_2^2 - \|v_i^{(0)}\|_2^2}{2} \\
&\overset{(d)}{>} \frac{(1+c_0-\gamma c_0)\log d}{d}\cdot\frac{\|w_i^{(T_0)}\|_2^2 + \|v_i^{(T_0)}\|_2^2}{2}.
\end{aligned} \tag{104}$$

In step (a), we use the definition of the sure-set $S_{1j,\text{sure}}$:

$$\left(\frac{\langle w_i^{(0)}, \mathbf{M}_{1j}\rangle + \langle v_i^{(0)}, \mathbf{M}_{1j}\rangle}{2}\right)^2 \geq \frac{(c_1+c_2)\log d}{d}\cdot\frac{\|\mathbf{M}_1\mathbf{M}_1^\top w_i^{(0)}\|_2^2 + \|\mathbf{M}_1\mathbf{M}_1^\top v_i^{(0)}\|_2^2}{4}. \tag{105}$$

In step (b), we apply Eq. (99). In step (c), we use the energy condition at T_0: the total energy in signal directions dominates the orthogonal components, combined with Eq. (101) and Eq. (103). In step (d), we use $c_1 + c_2 > 2(1 + c_0 - \gamma c_0)$ and the energy growth condition.

By symmetry of the coupled dynamics, the same bound holds for $|\langle v_i^{(T_0)}, \mathbf{M}_{1j}\rangle|^2$.

D.4. Upper Bound of Alignment for $i \notin S_{1j,\text{pot}}$

For neurons $i \notin S_{1j,\text{pot}}$, their weaker initialization results in smaller alignment:

$$\begin{aligned}
|\langle w_i^{(T_0)}, \mathbf{M}_{1j}\rangle|^2 &\overset{(a)}{\leq} (1+a_1)^{2T_0}\cdot\frac{(c_1-c_2)\log d}{d}\cdot\frac{\|\mathbf{M}_1\mathbf{M}_1^\top w_i^{(0)}\|_2^2 + \|\mathbf{M}_1\mathbf{M}_1^\top v_i^{(0)}\|_2^2}{4} \\
&= \frac{(c_1-c_2)\log d}{d}\cdot\frac{\|\mathbf{M}_1\mathbf{M}_1^\top w_i^{(T_0)}\|_2^2 + \|\mathbf{M}_1\mathbf{M}_1^\top v_i^{(T_0)}\|_2^2}{2} \\
&< \frac{(1+c_0-3\gamma c_0)\log d}{d}\cdot\frac{\|w_i^{(T_0)}\|_2^2 + \|v_i^{(T_0)}\|_2^2}{2}.
\end{aligned} \tag{106}$$

In step (a), we use the definition of the potential set $S_{1j,\text{pot}}$:

$$\left(\frac{\langle w_i^{(0)}, \mathbf{M}_{1j}\rangle + \langle v_i^{(0)}, \mathbf{M}_{1j}\rangle}{2}\right)^2 \leq \frac{(c_1-c_2)\log d}{d}\cdot\frac{\|\mathbf{M}_1\mathbf{M}_1^\top w_i^{(0)}\|_2^2 + \|\mathbf{M}_1\mathbf{M}_1^\top v_i^{(0)}\|_2^2}{4}. \tag{107}$$

By symmetry, the same bound holds for $|\langle v_i^{(T_0)}, \mathbf{M}_{1j}\rangle|^2$.

D.5. Weak Growth of $\mathbf{M}_2$ Alignment

For $\mathbf{M}_2$, all neurons exhibit weak growth due to the suppressed learning rate $a_2 = \Theta(\eta/d^{1+c_0})$. For any neuron $i \in [m]$ and any $j \in [d]$:

$$\begin{aligned}
|\langle w_i^{(T_0)}, \mathbf{M}_{2j}\rangle|^2 &= (1+a_2)^{2T_0}\left(\frac{\langle w_i^{(0)}, \mathbf{M}_{2j}\rangle + \langle v_i^{(0)}, \mathbf{M}_{2j}\rangle}{2}\right)^2 \\
&\leq \left(1+\frac{\eta}{d^{1+c_0}}\right)^{2T_0}\cdot O\left(\frac{\sigma_0^2}{d}\right) \\
&= O\left(\frac{\sigma_0^2}{d}\right),
\end{aligned} \tag{108}$$

where $\sigma_0^2 = \|w_i^{(0)}\|_2^2/d_1$ is the per-coordinate variance at initialization, and we used: (1) The initial projection satisfies $|\langle w_i^{(0)}, \mathbf{M}_{2j}\rangle|^2 = O(\sigma_0^2)$ by Gaussian concentration. (2) The growth factor satisfies $(1+a_2)^{T_0} = (1+\Theta(\eta/d^{1+c_0}))^{\Theta(d\log d/\eta)} = 1+o(1)$ when $c_0 > 0$. This confirms that $\mathbf{M}_2$ components remain at initialization scale throughout Stage 0.

D.6. Summary of Stage 0

At the end of Stage 0, the alignment properties are as follows:

- For neurons $i \in S_{1j,\mathrm{sure}}$:

$$|\langle w_i^{(T_0)}, \mathbf{M}_{1j}\rangle|^2 > \frac{(1+c_0-\gamma c_0)\log d}{d} \cdot \frac{\|w_i^{(T_0)}\|_2^2 + \|v_i^{(T_0)}\|_2^2}{2}. \tag{109}$$

- For neurons $i \notin S_{1j,\mathrm{pot}}$:

$$|\langle w_i^{(T_0)}, \mathbf{M}_{1j}\rangle|^2 < \frac{(1+c_0-3\gamma c_0)\log d}{d} \cdot \frac{\|w_i^{(T_0)}\|_2^2 + \|v_i^{(T_0)}\|_2^2}{2}. \tag{110}$$

- For $\mathbf{M}_2$:

$$|\langle w_i^{(T_0)}, \mathbf{M}_{2j}\rangle|^2 = O\left(\frac{\sigma_0^2}{d}\right). \tag{111}$$

- For orthogonal directions:

$$\|(I - \mathbf{M}_1\mathbf{M}_1^\top - \mathbf{M}_2\mathbf{M}_2^\top)w_i^{(T_0)}\|_2^2 < O\left(\frac{1}{d_1}\right) \cdot \|w_i^{(0)}\|_2^2. \tag{112}$$

These results demonstrate that during Stage 0: (1) Neurons in $S_{1j,\mathrm{sure}}$ develop significant alignment with $\mathbf{M}_{1j}$. (2) Neurons outside $S_{1j,\mathrm{pot}}$ remain weakly aligned with all $\mathbf{M}_1$ directions. (3) All neurons have negligible alignment with $\mathbf{M}_2$ directions (due to $a_1 \gg a_2$). (4) Orthogonal components remain at initialization scale.

The gap between $S_{1j,\mathrm{sure}}$ and $S_{1j,\mathrm{pot}}^c$ neurons is quantified by:

$$\frac{(1+c_0-\gamma c_0)-(1+c_0-3\gamma c_0)}{1+c_0} = \frac{2\gamma c_0}{1+c_0} > 0, \tag{113}$$

which provides a separation margin that will be exploited in Stage I.

E. Stage I: M_1 Alignment Phase

During Stage I when $T_0 < t \le T_1$ and $x_\tau = x_0 + \tau\epsilon$, the network predominantly aligns its parameters with $\mathbf{M}_1$. The $\mathbf{M}_2$ component remains suppressed due to the scale separation $\alpha_1 = \Theta(1)$ and $\alpha_2 = \Theta(1/d^{c_0})$.

Let the length of this training phase be denoted by $T_1 - T_0 = \Theta\left(\frac{d\log d}{\eta}\right)$. Throughout this stage, the projection of $w_i^{(t)}$ and $v_i^{(t)}$ onto $\mathbf{M}_1$ grows monotonically while the components along $\mathbf{M}_2$ remain at initialization scale.

At the beginning of this phase, we set the bias threshold as:

$$b_i^{(T_0)} = \sqrt{\frac{(1+c_0-2\gamma c_0)\log d}{d} \cdot \frac{\|w_i^{(T_0)}\|_2^2 + \|v_i^{(T_0)}\|_2^2}{2}}. \tag{114}$$

During training, the bias threshold is iteratively updated as:

$$b_i^{(t+1)} = \left(1+\frac{\eta}{d}\right) b_i^{(t)}, \tag{115}$$

for all $T_0 \le t < T_1$.

E.1. Alignment Growth for Neurons in $S_{1j,\text{sure}}$

This section analyzes the alignment behavior of neurons $i \in S_{1j,\text{sure}}$. For each $j \in [d]$, when $i \in S_{1j,\text{sure}}$, the projection $\langle w_i^{(t)}, \mathbf{M}_{1j} \rangle$ increases exponentially throughout Stage I ($T_0 < t \leq T_1$). This growth is driven by the signal along the $\mathbf{M}_1$ directions and the fact that the noise level in this stage does not disrupt the alignment dynamics.

E.1.1. Noise Projection Bound

For $i \in S_{1j,\text{sure}}$, using Lemma 1, the following holds with high probability $1 - e^{-\Omega(d_1)}$ when $T_0 < t \leq T_1$:

$$\left| \langle w_i^{(t)}, \tau\epsilon \rangle \right|^2 \leq O\left(\frac{\tau^2 \left\| w_i^{(t)} \right\|_2^2}{d^{1+c_0}} \right) < \left(b_i^{(t)} \right)^2. \tag{116}$$

E.1.2. Indicator Function Activation

Using the Stage 0 summary (Eq. 109) and the bias threshold (Eq. 114), we have at $t = T_0$:

$$|\langle w_i^{(T_0)}, \mathbf{M}_{1j} \rangle|^2 > \frac{(1 + c_0 - \gamma c_0) \log d}{d} \cdot \frac{\|w_i^{(T_0)}\|_2^2 + \|v_i^{(T_0)}\|_2^2}{2}. \tag{117}$$

Therefore, with high probability $1 - e^{-\Omega(d_1)}$, the indicator function satisfies the condition when $t = T_0$:

$$\mathbf{1}\{|w_i^{(T_0)\top} x_\tau| \geq b_i^{(T_0)}\} = 1 \quad \text{when } |z_{1j}| = 1. \tag{118}$$

We can ensure that:

$$\mathbb{E}\left[z_{1j}^2 \cdot \mathbf{1}\{w_i^{(t)\top} x_\tau \geq b_i^{(t)}\} \right] = \frac{C_z}{d}, \tag{119}$$

where $C_z = \Theta(1)$ is the sparsity constant with $C_z > 1$.

E.1.3. Alignment Growth Rate vs Bias Growth Rate

Using Eq. (115), we know that

$$\left(1 + \frac{\eta C_z}{d}\right) > \left(1 + \frac{\eta}{d}\right) \quad \text{since } C_z > 1. \tag{120}$$

Using the projected update dynamics (Eq. 125), we have:

$$|\langle w_i^{(t+1)}, \mathbf{M}_{1j} \rangle| > \left(1 + \frac{\eta}{d}\right) b_i^{(t)} = b_i^{(t+1)}. \tag{121}$$

This implies that when $t > T_0$, the alignment strength of informative features surpasses the updated bias threshold $b_i^{(t)}$. Consequently, the indicator functions become consistently activated for $T_0 < t \leq T_1$:

$$\mathbf{1}\{w_i^{(t)\top} x_\tau \geq b_i^{(t)}\} = 1 \quad \text{when } |z_{1j}| = 1. \tag{122}$$

E.1.4. Weight Dynamics

When the input satisfies $x_\tau = x_0 + \tau\epsilon$ and $|z_{1j}| = 1$, the SGD updates from the diffusion objective follow:

$$w_i^{(t+1)} = w_i^{(t)} + \eta \langle v_i^{(t)}, x_0 - g^{(t)}(x_\tau) \rangle x_\tau \, \mathbf{1}\{w_i^{(t)\top} x_\tau \geq b_i^{(t)}\}, \tag{123}$$

$$v_i^{(t+1)} = v_i^{(t)} + \eta \left(x_0 - g^{(t)}(x_\tau)\right) \left(w_i^{(t)\top} x_\tau\right) \mathbf{1}\{w_i^{(t)\top} x_\tau \geq b_i^{(t)}\}. \tag{124}$$

Projecting these updates onto the feature direction $\mathbf{M}_{1j}$:

$$\langle w_i^{(t+1)}, \mathbf{M}_{1j} \rangle = \langle w_i^{(t)}, \mathbf{M}_{1j} \rangle + \eta \langle v_i^{(t)}, \mathbf{M}_{1j} \rangle z_{1j} \, \mathbf{1}\{w_i^{(t)\top} x_\tau \geq b_i^{(t)}\} \pm \text{Err}_t, \tag{125}$$

$$\langle v_i^{(t+1)}, \mathbf{M}_{1j} \rangle = \langle v_i^{(t)}, \mathbf{M}_{1j} \rangle + \eta \, \langle w_i^{(t)}, \mathbf{M}_{1j} \rangle \, z_{1j} \, \mathbf{1}\{w_i^{(t)\top} x_\tau \geq b_i^{(t)}\} \pm \mathrm{Err}_t, \tag{126}$$

where Err_t collects contributions from the Gaussian perturbation $\langle \tau\epsilon, \mathbf{M}_{1j} \rangle$, the $\mathbf{M}_2$ component, and subdominant cross-terms inside $g^{(t)}(x_\tau)$.

Using Eq. (122), the weight dynamics can be expressed as:

$$\begin{bmatrix} \langle w_i^{(t+1)}, \mathbf{M}_{1j} \rangle \\ \langle v_i^{(t+1)}, \mathbf{M}_{1j} \rangle \end{bmatrix} = \begin{bmatrix} 1 & a_1 \\ a_1 & 1 \end{bmatrix} \begin{bmatrix} \langle w_i^{(t)}, \mathbf{M}_{1j} \rangle \\ \langle v_i^{(t)}, \mathbf{M}_{1j} \rangle \end{bmatrix} \pm \mathrm{Err}_t, \tag{127}$$

where $a_1 = \Theta\left(\frac{\eta}{d}\right)$.

Iterating from $t = T_0$:

$$\begin{aligned} \langle w_i^{(t)}, \mathbf{M}_{1j} \rangle &= \frac{(1+a_1)^{t-T_0} + (1-a_1)^{t-T_0}}{2} \langle w_i^{(T_0)}, \mathbf{M}_{1j} \rangle \\ &\quad + \frac{(1+a_1)^{t-T_0} - (1-a_1)^{t-T_0}}{2} \langle v_i^{(T_0)}, \mathbf{M}_{1j} \rangle \\ &= \left(1 + \Theta(\tfrac{\eta}{d})\right)^{t-T_0} \frac{\langle w_i^{(T_0)}, \mathbf{M}_{1j} \rangle + \langle v_i^{(T_0)}, \mathbf{M}_{1j} \rangle}{2}. \end{aligned} \tag{128}$$

Similarly for $v_i^{(t)}$:

$$\langle v_i^{(t)}, \mathbf{M}_{1j} \rangle = \left(1 + \Theta(\tfrac{\eta}{d})\right)^{t-T_0} \cdot \frac{\langle w_i^{(T_0)}, \mathbf{M}_{1j} \rangle + \langle v_i^{(T_0)}, \mathbf{M}_{1j} \rangle}{2}. \tag{129}$$

E.1.5. End of Stage I

At $t = T_1$, with $T_1 - T_0 = \Theta\left(\frac{d \log d}{\eta}\right)$, we have:

$$\left(1 + \Theta(\tfrac{\eta}{d})\right)^{T_1 - T_0} = \left(1 + \Theta(\tfrac{\eta}{d})\right)^{\Theta(d \log d/\eta)} = \mathrm{poly}(d). \tag{130}$$

Using the Stage 0 initial condition (Eq. 109):

$$\langle w_i^{(T_1)}, \mathbf{M}_{1j} \rangle \geq \mathrm{poly}(d) \cdot \sqrt{\frac{(1 + c_0 - \gamma c_0) \log d}{d} \cdot \frac{\|w_i^{(T_0)}\|_2^2 + \|v_i^{(T_0)}\|_2^2}{2}}. \tag{131}$$

Since the alignment grows exponentially while other directions grow at most polynomially:

$$\langle w_i^{(T_1)}, \mathbf{M}_{1j} \rangle \geq (1 - o(1)) \, \|w_i^{(T_1)}\|_2. \tag{132}$$

By symmetry:

$$\langle v_i^{(T_1)}, \mathbf{M}_{1j} \rangle \geq (1 - o(1)) \, \|v_i^{(T_1)}\|_2. \tag{133}$$

The training terminates Stage I when:

$$\|w_i^{(T_1)}\|_2^2 \geq \Omega(d) \|w_i^{(T_0)}\|_2^2. \tag{134}$$

E.2. Alignment for $i \notin S_{1j,\mathrm{sure}}$

This section examines the alignment behavior of neurons $i \notin S_{1j,\mathrm{sure}}$. For such neurons, the projection $\langle w_i^{(t)}, \mathbf{M}_{1j} \rangle$ remains small throughout Stage I ($T_0 < t \leq T_1$), exhibiting only negligible growth.

E.2.1. Indicator Function Deactivation

For $i \notin S_{1j,\text{sure}}$, using Eq. (116), Eq. (114), and the Stage 0 summary (Eq. 110), we have with high probability $1 - e^{-\Omega(d_1)}$ at $t = T_0$:

$$|\langle w_i^{(T_0)}, \mathbf{M}_{1j} \rangle|^2 < \frac{(1 + c_0 - 3\gamma c_0) \log d}{d} \cdot \frac{\|w_i^{(T_0)}\|_2^2 + \|v_i^{(T_0)}\|_2^2}{2} < \left(b_i^{(T_0)}\right)^2. \tag{135}$$

Therefore, the indicator function satisfies:

$$\mathbf{1}\{w_i^{(T_0)\top} x_\tau \geq b_i^{(T_0)}\} = 0. \tag{136}$$

We can ensure that:

$$\mathbb{E}\left[z_{1j}^2 \cdot \mathbf{1}\{w_i^{(t)\top} x_\tau \geq b_i^{(t)}\}\right] \leq o\left(\frac{1}{d^2}\right). \tag{137}$$

E.2.2. Alignment Growth Rate vs Bias Growth Rate

Using Eq. (115), we know that:

$$\left(1 + o\left(\frac{\eta}{d^2}\right)\right) < \left(1 + \frac{\eta}{d}\right). \tag{138}$$

Using the projected update dynamics, we have:

$$|\langle w_i^{(t+1)}, \mathbf{M}_{1j} \rangle| < \left(1 + \frac{\eta}{d}\right) b_i^{(t)} = b_i^{(t+1)}. \tag{139}$$

This implies that when $t > T_0$, the alignment strength does not surpass the updated bias threshold $b_i^{(t)}$. Consequently, the indicator functions remain consistently deactivated for $T_0 < t \leq T_1$:

$$\mathbf{1}\{w_i^{(t)\top} x_\tau \geq b_i^{(t)}\} = 0. \tag{140}$$

E.2.3. Suppressed Weight Dynamics

With the indicator deactivated, the weight dynamics become:

$$|\langle w_i^{(t+1)}, \mathbf{M}_{1j} \rangle| \leq \left(1 + o\left(\frac{\eta}{d^2}\right)\right)^{t - T_0} \left(\frac{\langle w_i^{(T_0)}, \mathbf{M}_{1j} \rangle + \langle v_i^{(T_0)}, \mathbf{M}_{1j} \rangle}{2}\right). \tag{141}$$

Because $\left(1 + o\left(\frac{\eta}{d^2}\right)\right)^{T_1 - T_0} \leq 1 + o\left(\frac{1}{d}\right)$, the growth in alignment is negligible:

$$|\langle w_i^{(T_1)}, \mathbf{M}_{1j} \rangle|^2 \leq \left(1 + o\left(\frac{1}{d}\right)\right) |\langle w_i^{(T_0)}, \mathbf{M}_{1j} \rangle|^2. \tag{142}$$

Using the Stage 0 bound (Eq. 110):

$$\langle w_i^{(T_1)}, \mathbf{M}_{1j} \rangle \leq O\left(\frac{1}{d^{1-c_0}}\right) \|w_i^{(T_1)}\|_2. \tag{143}$$

By symmetry:

$$\langle v_i^{(T_1)}, \mathbf{M}_{1j} \rangle \leq O\left(\frac{1}{d^{1-c_0}}\right) \|v_i^{(T_1)}\|_2. \tag{144}$$

E.3. Suppression of $\mathbf{M}_2$ Alignment in Stage I

We show that during Stage I, the $\mathbf{M}_2$ component cannot obtain nontrivial alignment for any neuron, because the bias threshold grows at rate $(1+\eta/d)$ while the $\mathbf{M}_2$-driven alignment (even under activation) grows only at rate $(1+\Theta(\eta/d^{1+c_0}))$, which is asymptotically slower.

Fix any neuron $i \in [m]$ and any coordinate $k \in [d]$. Consider samples where $z_{2k} = 1$. Decompose the pre-activation:

$$w_i^{(t)\top} x_\tau = \underbrace{\alpha_1 \langle w_i^{(t)}, \mathbf{M}_1 z_1 \rangle}_{\mathbf{M}_1 \text{ signal}} + \underbrace{\alpha_2 \langle w_i^{(t)}, \mathbf{M}_2 z_2 \rangle}_{\mathbf{M}_2 \text{ signal}} + \underbrace{\tau \langle w_i^{(t)}, \epsilon \rangle}_{\text{noise}}. \tag{145}$$

When we condition on $z_{2k} = 1$, the additional contribution that is uniquely attributable to coordinate k is at most

$$\alpha_2 \, |\langle w_i^{(t)}, \mathbf{M}_{2k} \rangle| \le \alpha_2 \, \|w_i^{(t)}\|_2 = \Theta(d^{-c_0}) \|w_i^{(t)}\|_2, \tag{146}$$

where we used $\|\mathbf{M}_{2k}\|_2 = 1$. In contrast, the bias threshold satisfies (by construction at T_0 and the update rule (115))

$$b_i^{(t)} = \Big(1 + \tfrac{\eta}{d}\Big)^{t - T_0} b_i^{(T_0)} \asymp \Big(1 + \tfrac{\eta}{d}\Big)^{t - T_0} \sqrt{\frac{\log d}{d}} \, \|w_i^{(T_0)}\|_2, \tag{147}$$

up to constant factors. Because $\Theta(d^{-c_0}) \gg \sqrt{\log d / d}$ may hold for small c_0, the only way $\mathbf{M}_2$ could activate the gate systematically is if $|\langle w_i^{(t)}, \mathbf{M}_{2k} \rangle|$ itself grew to be a non-negligible fraction of $\|w_i^{(t)}\|_2$. We now rule this out via a growth-rate comparison: even under optimistic always-on gating, $\mathbf{M}_2$ alignment cannot keep up with $b_i^{(t)}$.

Define the activation frequency for $\mathbf{M}_2$ coordinate k:

$$p_{i,k}^{(t)} \;=\; \mathbb{E}\Big[z_{2k}^2 \cdot \mathbf{1}\{ w_i^{(t)\top} x_\tau \ge b_i^{(t)} \} \Big]. \tag{148}$$

We will show $p_{i,k}^{(t)}$ is negligible throughout Stage I, which implies the $\mathbf{M}_2$-driven update is too rare to matter.

Projecting the SGD update onto $\mathbf{M}_{2k}$ and taking conditional expectation yields the analogue of (125)–(126):

$$\begin{aligned} \langle w_i^{(t+1)}, \mathbf{M}_{2k} \rangle &= \langle w_i^{(t)}, \mathbf{M}_{2k} \rangle + \eta \, \langle v_i^{(t)}, \mathbf{M}_{2k} \rangle \, \alpha_2 z_{2k} \, \mathbf{1}\{ w_i^{(t)\top} x_\tau \ge b_i^{(t)} \} \pm \mathrm{Err}_t, \\ \langle v_i^{(t+1)}, \mathbf{M}_{2k} \rangle &= \langle v_i^{(t)}, \mathbf{M}_{2k} \rangle + \eta \, \langle w_i^{(t)}, \mathbf{M}_{2k} \rangle \, \alpha_2 z_{2k} \, \mathbf{1}\{ w_i^{(t)\top} x_\tau \ge b_i^{(t)} \} \pm \mathrm{Err}_t, \end{aligned} \tag{149}$$

where Err_t collects contributions from $\tau\epsilon$, cross-terms, and the imperfect predictor $g^{(t)}(x_\tau)$ exactly as in the $\mathbf{M}_1$ analysis.

Taking absolute values and using $|z_{2k}| \le 1$, we obtain the growth envelope

$$\begin{bmatrix} |\langle w_i^{(t+1)}, \mathbf{M}_{2k} \rangle| \\ |\langle v_i^{(t+1)}, \mathbf{M}_{2k} \rangle| \end{bmatrix} \le \begin{bmatrix} 1 & \eta\alpha_2 \, \mathbf{1}\{ w_i^{(t)\top} x_\tau \ge b_i^{(t)} \} \\ \eta\alpha_2 \, \mathbf{1}\{ w_i^{(t)\top} x_\tau \ge b_i^{(t)} \} & 1 \end{bmatrix} \begin{bmatrix} |\langle w_i^{(t)}, \mathbf{M}_{2k} \rangle| \\ |\langle v_i^{(t)}, \mathbf{M}_{2k} \rangle| \end{bmatrix} + |\mathrm{Err}_t|. \tag{150}$$

If we over-approximate the indicator by its mean frequency $p_{i,k}^{(t)}$ and ignore Err_t, the effective coupling coefficient is

$$a_2^{\mathrm{eff}} = \eta\alpha_2 \, \frac{p_{i,k}^{(t)}}{\mathbb{E}[z_{2k}^2]} = \eta\alpha_2 \, d \, p_{i,k}^{(t)} \le \eta\alpha_2 \, d \cdot \mathbb{E}[z_{2k}^2] = \Theta\Big(\frac{\eta}{d^{1+c_0}} \Big), \tag{151}$$

since $\mathbb{E}[z_{2k}^2] = \Theta(1/d)$. Thus, even in the best-case scenario where the gate is always active whenever $z_{2k} = 1$, the multiplicative growth per step along $\mathbf{M}_{2k}$ is at most

$$1 + \Theta\Big(\frac{\eta}{d^{1+c_0}} \Big). \tag{152}$$

Meanwhile the bias grows per step as $1 + \eta/d$ by (115). Since $c_0 > 0$,

$$1 + \Theta\Big(\frac{\eta}{d^{1+c_0}} \Big) < 1 + \frac{\eta}{d}. \tag{153}$$

Therefore, the ratio

$$R_{i,k}^{(t)} \;=\; \frac{|\langle w_i^{(t)}, \mathbf{M}_{2k} \rangle|}{b_i^{(t)}} \tag{154}$$

cannot increase over Stage I (up to the negligible Err_t terms controlled by Lemma 1 and the same orthogonal-growth bounds used in Stage 0). Combining (152) with (115) gives the multiplicative drift

$$R_{i,k}^{(t+1)} \le \frac{1+\Theta(\eta/d^{1+c_0})}{1+\eta/d} R_{i,k}^{(t)} \le \Big(1-\Theta(d^{-c_0})\Big) R_{i,k}^{(t)}. \tag{155}$$

Hence, if $R_{i,k}^{(T_0)} < 1$ (which holds w.h.p. from the Stage 0 bound (111) together with the choice of $b_i^{(T_0)}$), then $R_{i,k}^{(t)} < 1$ for all $T_0 < t \le T_1$, and the gate cannot become consistently active due to $\mathbf{M}_2$.

Since $\mathbf{M}_2$ cannot sustain activation and its effective coupling is at most $\Theta(\eta/d^{1+c_0})$, iterating (149) over $T_1 - T_0 = \Theta(d\log d/\eta)$ steps yields

$$|\langle w_i^{(T_1)}, \mathbf{M}_{2k}\rangle|^2 \le \Big(1+\Theta(\tfrac{\eta}{d^{1+c_0}})\Big)^{2(T_1-T_0)} |\langle w_i^{(T_0)}, \mathbf{M}_{2k}\rangle|^2 + \sum_{t=T_0}^{T_1-1} |\mathrm{Err}_t|^2 = (1+o(1))|\langle w_i^{(T_0)}, \mathbf{M}_{2k}\rangle|^2 + o\Big(\|w_i^{(T_1)}\|_2^2\Big). \tag{156}$$

Using the Stage 0 initialization-scale bound (111), we conclude that for all $i \in [m]$ and $k \in [d]$,

$$|\langle w_i^{(T_1)}, \mathbf{M}_{2k}\rangle|^2 = O\left(\frac{1}{d^{2c_0}}\right)\|w_i^{(T_1)}\|_2^2, \qquad |\langle v_i^{(T_1)}, \mathbf{M}_{2k}\rangle|^2 = O\left(\frac{1}{d^{2c_0}}\right)\|v_i^{(T_1)}\|_2^2, \tag{157}$$

which matches the claimed Stage I summary scaling and formalizes the suppression mechanism: $\mathbf{M}_2$ is too weak to overcome the growing bias threshold, and even optimistic activation cannot compensate for the α_2-suppressed learning rate.

E.4. Summary of Stage I

At the end of Stage I ($t = T_1$), with $T_1 - T_0 = \Theta\left(\frac{d\log d}{\eta}\right)$, the alignment properties are characterized as follows.

- For $i \in S_{1j,\mathrm{sure}}$:
$$|\langle w_i^{(T_1)}, \mathbf{M}_{1j}\rangle|^2 \ge (1-o(1))\|w_i^{(T_1)}\|_2^2. \tag{158}$$
- For $i \notin S_{1j,\mathrm{sure}}$:
$$|\langle w_i^{(T_1)}, \mathbf{M}_{1j}\rangle|^2 \le O\left(\frac{\log d}{d}\right)\cdot\|w_i^{(T_1)}\|_2^2. \tag{159}$$
- For all neurons and $\mathbf{M}_2$:
$$|\langle w_i^{(T_1)}, \mathbf{M}_{2k}\rangle|^2 = O\left(\frac{1}{d^{2c_0}}\right)\cdot\|w_i^{(T_1)}\|_2^2 \tag{160}$$
for all $i \in [m]$ and $k \in [d]$.
- For orthogonal directions:
$$\|(I - \mathbf{M}_1\mathbf{M}_1^\top - \mathbf{M}_2\mathbf{M}_2^\top)w_i^{(T_1)}\|_2^2 \le O\left(\frac{1}{d_1}\right)\cdot\|w_i^{(T_0)}\|_2^2. \tag{161}$$

These results demonstrate that during Stage I: (1) Neurons in $S_{1j,\mathrm{sure}}$ achieve near-perfect alignment with $\mathbf{M}_{1j}$. (2) Neurons outside $S_{1j,\mathrm{sure}}$ remain weakly aligned with all $\mathbf{M}_1$ directions. (3) All neurons have negligible alignment with $\mathbf{M}_2$ directions, since $\alpha_2 = \Theta(1/d^{c_0})$ is insufficient to overcome the bias threshold. (4) Orthogonal components remain at initialization scale.

The key mechanism is the growth rate hierarchy:

$$\underbrace{\left(1+\frac{\eta C_z}{d}\right)^{T_1-T_0}}_{\mathbf{M}_1 \text{ for } i\in S_{1j,\mathrm{sure}}} = \mathrm{poly}(d) \gg \underbrace{\left(1+\frac{\eta}{d^{1+2c_0}}\right)^{T_1-T_0}}_{\mathbf{M}_2 \text{ for all } i} = 1+o(1). \tag{162}$$

F. Stage II: M_2 Alignment Phase (High-to-Low Denoising Schedule Only)

F.1. Setup and Goal

Under the **high-to-low denoising schedule only** protocol: in Stage I, the noise level is sampled from $\tau \sim \mathrm{Unif}[\tau_1, \tau_{\max}]$, and in Stage II, the noise level is sampled from $\tau \sim \mathrm{Unif}[\tau_{\min}, \tau_1]$, where

$$\tau_{\min} = \Theta(1/d_1), \quad \tau_1 = \Theta(\alpha_2/\sqrt{\log d}), \quad \tau_{\max} = \Theta(\alpha_1/\sqrt{\log d}).$$

Crucially, **no neurons are frozen** during Stage I, so all neurons participate in training and accumulate non-negligible $\mathbf{M}_1$ alignment before Stage II begins. Since $\tau_1 = \Theta(\alpha_2/\sqrt{\log d}) < \alpha_2$, the noise level in Stage II is smaller than the $\mathbf{M}_2$ signal strength, enabling $\mathbf{M}_2$ feature recovery.

In this section, we prove that for neurons $i \in S_{2j,\mathrm{sure}}$, the coordinate $\langle w_i^{(t)}, \mathbf{M}_{2j}\rangle$ first increases slowly (because the gate is mainly triggered by $\mathbf{M}_1$-alignment), and after it crosses the bias threshold, the $\mathbf{M}_{2j}$ direction becomes self-sustaining and enters an exponential-growth regime driven by the $\mathbf{M}_2$ signal.

Throughout Stage II, we keep the same bias update rule as Stage I:

$$b_i^{(t+1)} = \Big(1 + \frac{\eta}{d}\Big)\, b_i^{(t)}. \tag{163}$$

For $i \in S_{2j,\mathrm{sure}}$, we show that there exists a hitting time $t_\star \in [T_1, T_1 + T_2]$ such that

$$|\langle w_i^{(t_\star)}, \mathbf{M}_{2j}\rangle| \geq b_i^{(t_\star)}, \tag{164}$$

and for all $t \geq t_\star$, the gate stays active whenever $z_{2j} = 1$, yielding

$$|\langle w_i^{(T_1+T_2)}, \mathbf{M}_{2j}\rangle| \geq \left(1 + \Theta\Big(\frac{\alpha_2^2\eta}{d}\Big)\right)^{T_2-(t_\star-T_1)} \cdot \left|\frac{\langle w_i^{(t_\star)}, \mathbf{M}_{2j}\rangle + \langle v_i^{(t_\star)}, \mathbf{M}_{2j}\rangle}{2}\right| - o(1). \tag{165}$$

F.2. Projected Dynamics Along $\mathbf{M}_{2j}$

We analyze the projected SGD updates onto $\mathbf{M}_{2j}$. Using the same projection argument as in Stage I, we obtain the coupled recursion:

$$\begin{aligned}
\langle w_i^{(t+1)}, \mathbf{M}_{2j}\rangle &= \langle w_i^{(t)}, \mathbf{M}_{2j}\rangle + \eta\, \langle v_i^{(t)}, \mathbf{M}_{2j}\rangle\, \alpha_2^2 z_{2j}^2\, \mathbf{1}\{w_i^{(t)\top} x_\tau \geq b_i^{(t)}\} \pm \mathrm{Err}_t, \\
\langle v_i^{(t+1)}, \mathbf{M}_{2j}\rangle &= \langle v_i^{(t)}, \mathbf{M}_{2j}\rangle + \eta\, \langle w_i^{(t)}, \mathbf{M}_{2j}\rangle\, \alpha_2^2 z_{2j}^2\, \mathbf{1}\{w_i^{(t)\top} x_\tau \geq b_i^{(t)}\} \pm \mathrm{Err}_t.
\end{aligned} \tag{166}$$

Here Err_t collects all non-leading contributions, including: (i) the Gaussian perturbation term $\tau\epsilon$, (ii) cross-terms between $\mathbf{M}_1$ and $\mathbf{M}_2$, (iii) subdominant coordinates $\mathbf{M}_{2j'}$ for $j' \neq j$, (iv) approximation errors from $g^{(t)}(x_\tau)$. We assume the same concentration control as in Stage I: with probability $1 - e^{-\Omega(d_1)}$,

$$|\mathrm{Err}_t| \leq \frac{1}{\mathrm{poly}(d)}\Big(|\langle w_i^{(t)}, \mathbf{M}_{2j}\rangle| + |\langle v_i^{(t)}, \mathbf{M}_{2j}\rangle| + \|w_i^{(t)}\|_2 + \|v_i^{(t)}\|_2\Big), \tag{167}$$

uniformly for all $t \in [T_1, T_1 + T_2]$. This is a direct analogue of the noise projection bounds used in Stage 0–I, and can be proved using Gaussian concentration and the fact that $\mathbf{M}_2$ components remain small before activation.

Define the gate frequency conditioned on $z_{2j} = 1$:

$$p_{i,j}^{(t)} = \Pr\Big(w_i^{(t)\top} x_\tau \geq b_i^{(t)} \,\Big|\, z_{2j} = 1\Big). \tag{168}$$

Taking conditional expectation of (166) given $z_{2j} = 1$ yields

$$\mathbb{E}\Big[\alpha_2^2 z_{2j}^2 \mathbf{1}\{w_i^{(t)\top} x_\tau \geq b_i^{(t)}\}\Big] = \Theta\Big(\frac{\alpha_2^2}{d}\Big) \cdot p_{i,j}^{(t)}. \tag{169}$$

Thus the effective coupling coefficient is

$$a_{2,t}^{\mathrm{eff}} = \Theta\Big(\frac{\alpha_2^2\eta}{d}\Big) \cdot p_{i,j}^{(t)}. \tag{170}$$

F.3. Regime A: Slow Growth Before Hitting the Bias

In the beginning of Stage II, $\mathbf{M}_2$ components are still at initialization scale. In this regime, the gate is mainly triggered by the already-established $\mathbf{M}_1$ alignment. We formalize this by assuming there exists a constant $c_{\mathrm{gate}} \in (0, 1]$ such that

$$p_{i,j}^{(t)} \geq c_{\mathrm{gate}} \quad \text{for all } t \in [T_1, t_\star), \tag{171}$$

meaning the neuron fires a constant fraction of the time due to $\mathbf{M}_1$-driven activation.

Combining (170) and (171), we obtain a uniform lower bound on the effective coupling:

$$a_{2,t}^{\mathrm{eff}} \geq \Theta\left(\frac{\alpha_2^2 \eta}{d}\right). \tag{172}$$

Ignoring Err_t and iterating the symmetric 2D system as in Stage 0–I yields

$$\left|\frac{\langle w_i^{(t)}, \mathbf{M}_{2j}\rangle + \langle v_i^{(t)}, \mathbf{M}_{2j}\rangle}{2}\right| \geq \left(1 + \Theta\left(\frac{\alpha_2^2 \eta}{d}\right)\right)^{t-T_1} \cdot \left|\frac{\langle w_i^{(T_1)}, \mathbf{M}_{2j}\rangle + \langle v_i^{(T_1)}, \mathbf{M}_{2j}\rangle}{2}\right| - \sum_{s=T_1}^{t-1} |\mathrm{Err}_s|. \tag{173}$$

By Stage I suppression, we have initialization-scale seed:

$$\left|\frac{\langle w_i^{(T_1)}, \mathbf{M}_{2j}\rangle + \langle v_i^{(T_1)}, \mathbf{M}_{2j}\rangle}{2}\right| = \Theta\left(\frac{1}{\sqrt{d}}\right) \quad \text{w.h.p.} \tag{174}$$

Combining (174) with (167) implies the error sum is negligible relative to the main term for all $t \leq T_1 + T_2$, so the slow growth is dominated by the factor $\left(1 + \Theta(\alpha_2^2 \eta/d)\right)^{t-T_1}$.

F.4. Hitting Time and Definition of $S_{2j,\mathrm{sure}}$

We define the sure set $S_{2j,\mathrm{sure}}$ so that for each $i \in S_{2j,\mathrm{sure}}$, the coordinate j is guaranteed to be the first $\mathbf{M}_2$-direction that crosses the bias: there exists $t_\star \in [T_1, T_1 + T_2]$ such that (164) holds and for all $j' \neq j$,

$$|\langle w_i^{(t)}, \mathbf{M}_{2j'}\rangle| < b_i^{(t)} \quad \text{for all } t \leq t_\star. \tag{175}$$

Using (173), the hitting condition (164) is satisfied provided

$$\left(1 + \Theta\left(\frac{\alpha_2^2 \eta}{d}\right)\right)^{t_\star - T_1} \cdot \Theta\left(\frac{1}{\sqrt{d}}\right) \geq b_i^{(t_\star)}. \tag{176}$$

Since the bias evolves by (163), we have

$$b_i^{(t_\star)} = \left(1 + \frac{\eta}{d}\right)^{t_\star - T_1} b_i^{(T_1)}. \tag{177}$$

Thus (176) holds when Stage II is long enough so that the $\mathbf{M}_2$ amplification overcomes the bias scaling. A sufficient choice is

$$T_2 = \Theta\left(\frac{d}{\alpha_2^2} \cdot \frac{\log d}{\eta}\right) = \Theta\left(\frac{d^{1+2c_0} \log d}{\eta}\right), \tag{178}$$

because then

$$\left(1 + \Theta\left(\frac{\alpha_2^2 \eta}{d}\right)\right)^{T_2} = \mathrm{poly}(d), \tag{179}$$

while $(1 + \eta/d)^{T_2}$ is also $\mathrm{poly}(d)$, and the sure-set condition selects those neurons for which the net ratio in (176) crosses 1.

F.5. Regime B: Exponential Growth After Crossing the Bias

We now analyze the post-threshold dynamics. Fix a neuron $i \in S_{2j,\mathrm{sure}}$, and let $t_\star \in [T_1, T_1 + T_2]$ denote the first time such that the alignment along $\mathbf{M}_{2j}$ reaches the activation scale. By the analysis in Regime A, this event is characterized by the relative alignment condition

$$|\langle w_i^{(t_\star)}, \mathbf{M}_{2j}\rangle| \;\geq\; \Theta\left(\sqrt{\frac{\log d}{d}}\right) \|w_i^{(t_\star)}\|_2, \tag{180}$$

with high probability. At this point, the coordinate $\mathbf{M}_{2j}$ becomes the dominant contributor to the pre-activation among all $\mathbf{M}_2$ directions.

Conditioned on $z_{2j} = 1$, the contribution of $\mathbf{M}_{2j}$ to $w_i^{(t)\top} x_\tau$ exceeds the noise and all other $\mathbf{M}_2$ coordinates by a polynomial margin. By Lemma 4 and the definition of the sure set $S_{2j,\mathrm{sure}}$, this implies that for all $t \geq t_\star$,

$$\mathbf{1}\{w_i^{(t)\top} x_\tau \geq b_i^{(t)}\} = 1 \quad \text{whenever } z_{2j} = 1, \tag{181}$$

with probability at least $1 - e^{-\Omega(d_1)}$. In particular, the activation frequency conditioned on $z_{2j} = 1$ satisfies

$$\Pr\Big(w_i^{(t)\top} x_\tau \geq b_i^{(t)} \mid z_{2j} = 1\Big) = 1 - o(1), \qquad \forall t \geq t_\star. \tag{182}$$

Crucially, for any $j' \neq j$, the contribution from $\mathbf{M}_{2j'}$ remains below the bias threshold, so by Lemma 4 (Regime (A)):

$$\Pr\Big(w_i^{(t)\top} x_\tau \geq b_i^{(t)} \mid z_{2j'} = 1\Big) = o(1), \qquad \forall j' \neq j,\ \forall t \geq t_\star. \tag{183}$$

Equations (182) and (183) together establish **gating separation**: the neuron almost always activates when the correct weak feature $z_{2j} = 1$ is present, and almost never fires for any other weak feature direction $z_{2j'} = 1$ with $j' \neq j$.

Substituting (182) into the projected update equations (166), the dynamics along $\mathbf{M}_{2j}$ reduce, for all $t \geq t_\star$, to the symmetric linear system

$$\begin{bmatrix} \langle w_i^{(t+1)}, \mathbf{M}_{2j}\rangle \\ \langle v_i^{(t+1)}, \mathbf{M}_{2j}\rangle \end{bmatrix} = \begin{bmatrix} 1 & \tilde{a} \\ \tilde{a} & 1 \end{bmatrix} \begin{bmatrix} \langle w_i^{(t)}, \mathbf{M}_{2j}\rangle \\ \langle v_i^{(t)}, \mathbf{M}_{2j}\rangle \end{bmatrix} \pm \mathrm{Err}_t, \qquad \tilde{a} = \Theta\left(\frac{\alpha_2^2 \eta}{d}\right), \tag{184}$$

where Err_t satisfies the uniform bound in (167).

Ignoring the negligible error term and diagonalizing the update matrix in (184), we obtain for all $t \geq t_\star$,

$$|\langle w_i^{(t)}, \mathbf{M}_{2j}\rangle| \;\geq\; \left(1 + \Theta\left(\frac{\alpha_2^2 \eta}{d}\right)\right)^{t - t_\star} \cdot |\langle w_i^{(t_\star)}, \mathbf{M}_{2j}\rangle| - o\Big(\|w_i^{(t)}\|_2\Big), \tag{185}$$

and the same bound holds for $|\langle v_i^{(t)}, \mathbf{M}_{2j}\rangle|$ by symmetry of the coupled dynamics.

Combining (185) with the crossing condition (180), we conclude that the $\mathbf{M}_{2j}$ alignment grows exponentially from time $t_\star$ onward, at rate $1 + \Theta(\alpha_2^2 \eta / d)$.

Since the $\mathbf{M}_1$ components remain stable during Stage II and all other $\mathbf{M}_2$ coordinates are suppressed by the sure-set condition, the norm of $w_i^{(t)}$ is asymptotically dominated by the $\mathbf{M}_{2j}$ component for t sufficiently larger than $t_\star$. In particular, for $t = T_1 + T_2$,

$$|\langle w_i^{(T_1+T_2)}, \mathbf{M}_{2j}\rangle| = \Theta(1)\|w_i^{(T_1+T_2)}\|_2 \tag{186}$$

and analogously for $v_i^{(T_1+T_2)}$.

F.6. Stability of $\mathbf{M}_1$ Components During Stage II

We show that throughout Stage II, the components of the network parameters along $\mathbf{M}_1$ remain stable, up to negligible lower-order fluctuations. This holds uniformly for all neurons $i \in [m]$, regardless of whether $i \in S_{1j,\mathrm{sure}}$ or $i \notin S_{1j,\mathrm{sure}}$.

The key reason is that by the end of Stage I, the predictor $g^{(t)}$ has already learned the $\mathbf{M}_1$ structure to high accuracy. As a result, the residual signal driving further updates along $\mathbf{M}_1$ directions is uniformly small during Stage II.

For any atom $\mathbf{M}_{1j'}$, the residual projection satisfies

$$\left|\langle x_\tau - g^{(t)}(x_\tau), \mathbf{M}_{1j'}\rangle\right| \le O\left(\frac{1}{d}\right), \qquad \forall t \in [T_1, T_1+T_2], \tag{187}$$

with high probability. This follows from the Stage I analysis, which establishes that the dominant $\mathbf{M}_1$ components have already been accurately captured by the model, leaving only vanishing residuals in these directions.

Consider the projected update of $w_i^{(t)}$ along any $\mathbf{M}_{1j'}$. Using (187) and the SGD update rule, we obtain

$$\mathbb{E}\Big[\langle v_i^{(t)}, x_\tau - g^{(t)}(x_\tau)\rangle\, z_{1j'}\, \mathbf{1}\{w_i^{(t)\top} x_\tau \ge b_i^{(t)}\}\Big] \le O\left(\frac{1}{d^2}\right), \tag{188}$$

uniformly over all neurons i and all indices j'.

In addition, when $z_{1j'} = 0$, the pre-activation $\langle w_i^{(t)}, x_\tau\rangle$ is dominated by the Gaussian noise term and satisfies

$$\Pr\Big(\langle w_i^{(t)}, x_\tau\rangle \ge b_i^{(t)} \;\Big|\; z_{1j'} = 0\Big) \le O\left(\frac{1}{d^2}\right), \tag{189}$$

where we used Lemma 4 and the fact that $\tau = O(1/d)$ in Stage II. Therefore, both the magnitude of the gradient and the frequency of activation along $\mathbf{M}_1$ directions are strongly suppressed.

Combining the above bounds, we conclude that for any neuron $i \in [m]$ and any $\mathbf{M}_{1j'}$,

$$\left|\langle w_i^{(t+1)}, \mathbf{M}_{1j'}\rangle - \langle w_i^{(t)}, \mathbf{M}_{1j'}\rangle\right| \le \frac{1}{\mathrm{poly}(d)}\, \|w_i^{(t)}\|_2, \qquad \forall t \in [T_1, T_1+T_2], \tag{190}$$

and an analogous bound holds for $v_i^{(t)}$. Summing over t and using $T_2 = \mathrm{poly}(d)$, the total variation of the $\mathbf{M}_1$ components during Stage II is negligible compared to their Stage I magnitudes.

F.7. Summary of Stage II

At the end of Stage II ($t = T_1 + T_2$), with $T_2 = \Theta\Big(\frac{d^{1+2c_0}\log d}{\eta}\Big)$, the alignment properties are summarized as follows.

- For $i \in S_{1j,\mathrm{sure}}$: the $\mathbf{M}_1$ alignment learned in Stage I remains stable. In particular,

$$|\langle w_i^{(T_1+T_2)}, \mathbf{M}_{1j}\rangle|^2 \ge (1 - o(1))\, \|w_i^{(T_1+T_2)}\|_2^2, \tag{191}$$

  and all other $\mathbf{M}_1$ and $\mathbf{M}_2$ coordinates remain negligible. Thus, neurons in $S_{1j,\mathrm{sure}}$ preserve essentially pure $\mathbf{M}_{1j}$ representations throughout Stage II.

- For $i \notin S_{2j,\mathrm{sure}}$: the $\mathbf{M}_2$ components fail to cross the activation threshold, and their alignment along all $\mathbf{M}_2$ directions remains suppressed:

$$|\langle w_i^{(T_1+T_2)}, \mathbf{M}_{2k}\rangle|^2 \le O\left(\frac{1}{d^{2c_0}}\right)\, \|w_i^{(T_1+T_2)}\|_2^2, \qquad \forall k \in [d]. \tag{192}$$

  Their $\mathbf{M}_1$ components remain at the Stage I scale.

- For $i \in S_{2j,\mathrm{sure}}$: the neuron enters the exponential-growth regime along $\mathbf{M}_{2j}$ during Stage II, while the previously learned $\mathbf{M}_1$ alignment remains stable. Consequently, both components contribute at comparable scale:

$$|\langle w_i^{(T_1+T_2)}, \mathbf{M}_{1j}\rangle|^2 = \Theta(1)\, \|w_i^{(T_1+T_2)}\|_2^2, \qquad |\langle w_i^{(T_1+T_2)}, \mathbf{M}_{2j}\rangle|^2 = \Theta(1)\, \|w_i^{(T_1+T_2)}\|_2^2. \tag{193}$$

  In particular, neither the $\mathbf{M}_{1j}$ nor the $\mathbf{M}_{2j}$ coordinate dominates the norm, and the learned representation corresponds to a superposition of both dictionary components.

- For orthogonal directions: for all neurons $i \in [m]$,

$$\|(I - \mathbf{M}_1\mathbf{M}_1^\top - \mathbf{M}_2\mathbf{M}_2^\top) w_i^{(T_1+T_2)}\|_2^2 \le O\left(\frac{1}{d_1}\right)\, \|w_i^{(T_0)}\|_2^2. \tag{194}$$

Stage II does not produce pure $\mathbf{M}_2$ neurons. Instead, for neurons in $S_{2j,\mathrm{sure}}$, the dynamics lead to a stable coexistence of the previously learned coarse feature $\mathbf{M}_{1j}$ and the newly amplified fine feature $\mathbf{M}_{2j}$. As a result, the final representation cannot be attributed to a single dictionary atom but corresponds to an impure superposition across feature scales. Combined with the stable $\mathbf{M}_1$-aligned neurons from Stage I, the final network contains specialized neurons for both $\mathbf{M}_1$ and $\mathbf{M}_2$ directions.

G. Stage II: M_2 Alignment Phase (Joint Denoising-Sparsity Scheduling)

G.1. Setup and Goal

Under the **joint denoising-sparsity scheduling** protocol: in Stage I, the noise level is sampled from $\tau \sim \mathrm{Unif}[\tau_1, \tau_{\max}]$, and in Stage II, the noise level is sampled from $\tau \sim \mathrm{Unif}[\tau_{\min}, \tau_1]$, with the same noise thresholds as the denoise-only case:

$$\tau_{\min} = \Theta(1/d_1), \quad \tau_1 = \Theta(\alpha_2/\sqrt{\log d}), \quad \tau_{\max} = \Theta(\alpha_1/\sqrt{\log d}).$$

The crucial difference is the **model sparsity schedule**: during Stage I, we randomly select a subset $S_1 \subset [m]$ with $|S_1| = O(d^{1.01})$, and only neurons in S_1 are updated; the remaining neurons are **frozen** at their initialization. By construction, $S_{2j,\mathrm{sure}} \cap S_1 = \emptyset$ for all $j \in [d]$, meaning the $\mathbf{M}_2$-targeted neurons do not participate in Stage I training.

In this section, we prove that for neurons $i \in S_{2j,\mathrm{sure}}$, the coordinate $\langle \hat{w}_i^{(t)}, \mathbf{M}_{2j} \rangle$ grows exponentially during Stage II and achieves **pure** $\mathbf{M}_{2j}$ alignment, in contrast to the impure alignment observed under the denoise-only protocol.

Throughout Stage II, we use the same bias update rule:

$$\hat{b}_i^{(t+1)} = \left(1 + \frac{\eta}{d}\right) \hat{b}_i^{(t)}. \tag{195}$$

For $i \in S_{2j,\mathrm{sure}}$, we show that there exists a hitting time $t_\star \in [T_1, T_1 + T_2]$ such that

$$|\langle \hat{w}_i^{(t_\star)}, \mathbf{M}_{2j} \rangle| \geq \hat{b}_i^{(t_\star)}, \tag{196}$$

and for all $t \geq t_\star$, the gate stays active whenever $z_{2j} = 1$, yielding

$$|\langle \hat{w}_i^{(T_1+T_2)}, \mathbf{M}_{2j} \rangle|^2 = (1 - o(1)) \, \|\hat{w}_i^{(T_1+T_2)}\|_2^2. \tag{197}$$

G.2. Frozen State at End of Stage I

Since neurons in $S_{2j,\mathrm{sure}}$ are frozen during Stage I, their weights remain at the initialization scale:

$$\hat{w}_i^{(T_1)} = \hat{w}_i^{(0)}, \qquad \hat{v}_i^{(T_1)} = \hat{v}_i^{(0)}, \qquad \forall i \in S_{2j,\mathrm{sure}}. \tag{198}$$

By the initialization distribution $\hat{w}_i^{(0)}, \hat{v}_i^{(0)} \sim \mathcal{N}(0, \sigma_0^2 I_{d_1})$, the projections onto the dictionary atoms satisfy:

$$|\langle \hat{w}_i^{(T_1)}, \mathbf{M}_{1j'} \rangle| = O\left(\frac{\sigma_0}{\sqrt{d}}\right), \qquad \forall j' \in [d], \tag{199}$$

$$|\langle \hat{w}_i^{(T_1)}, \mathbf{M}_{2j} \rangle| = \Theta\left(\frac{\sigma_0}{\sqrt{d}}\right), \tag{200}$$

with high probability $1 - e^{-\Omega(d)}$.

The key observation is that unlike the denoise-only case where neurons accumulate $\mathbf{M}_1$ alignment during Stage I, here the $\mathbf{M}_1$ components remain at the initialization scale $O(\sigma_0/\sqrt{d})$, which is negligible compared to the final norm.

G.3. Projected Dynamics Along $\mathbf{M}_{2j}$

We analyze the projected SGD updates onto $\mathbf{M}_{2j}$ during Stage II. Using the same projection argument as in the denoise-only case, we obtain the coupled recursion:

$$\begin{aligned}
\langle \hat{w}_i^{(t+1)}, \mathbf{M}_{2j} \rangle &= \langle \hat{w}_i^{(t)}, \mathbf{M}_{2j} \rangle + \eta \, \langle \hat{v}_i^{(t)}, \mathbf{M}_{2j} \rangle \, \alpha_2^2 z_{2j}^2 \, \mathbf{1}\{\hat{w}_i^{(t)\top} x_\tau \geq \hat{b}_i^{(t)}\} \pm \widehat{\mathrm{Err}}_t, \\
\langle \hat{v}_i^{(t+1)}, \mathbf{M}_{2j} \rangle &= \langle \hat{v}_i^{(t)}, \mathbf{M}_{2j} \rangle + \eta \, \langle \hat{w}_i^{(t)}, \mathbf{M}_{2j} \rangle \, \alpha_2^2 z_{2j}^2 \, \mathbf{1}\{\hat{w}_i^{(t)\top} x_\tau \geq \hat{b}_i^{(t)}\} \pm \widehat{\mathrm{Err}}_t.
\end{aligned} \tag{201}$$

The error term $\widehat{\mathrm{Err}}_t$ collects contributions from: (i) the Gaussian perturbation term $\tau\epsilon$, (ii) cross-terms between $\mathbf{M}_1$ and $\mathbf{M}_2$, (iii) subdominant coordinates $\mathbf{M}_{2j'}$ for $j' \neq j$, (iv) approximation errors from $g^{(t)}(x_\tau)$. With probability $1 - e^{-\Omega(d_1)}$,

$$|\widehat{\mathrm{Err}}_t| \le \frac{1}{\mathrm{poly}(d)}\Big(|\langle \hat{w}_i^{(t)}, \mathbf{M}_{2j}\rangle| + |\langle \hat{v}_i^{(t)}, \mathbf{M}_{2j}\rangle| + \|\hat{w}_i^{(t)}\|_2 + \|\hat{v}_i^{(t)}\|_2\Big), \tag{202}$$

uniformly for all $t \in [T_1, T_1 + T_2]$.

G.4. Regime A: Slow Growth Before Hitting the Bias

At the beginning of Stage II, all components of $\hat{w}_i^{(t)}$ are at initialization scale. Unlike the denoise-only case where the gate is triggered by $\mathbf{M}_1$ alignment, here the gate activation is driven by the Gaussian component of the pre-activation.

Define the gate frequency conditioned on $z_{2j} = 1$:

$$\hat{p}_{i,j}^{(t)} = \Pr\Big(\hat{w}_i^{(t)\top} x_\tau \ge \hat{b}_i^{(t)} \,\Big|\, z_{2j} = 1\Big). \tag{203}$$

Since $\hat{w}_i^{(t)}$ is approximately Gaussian with norm $\|\hat{w}_i^{(t)}\|_2 = \Theta(\sigma_0\sqrt{d_1})$ and the bias $\hat{b}_i^{(t)} = O(\sigma_0\sqrt{\log d})$, by standard Gaussian tail bounds,

$$\hat{p}_{i,j}^{(t)} \ge c_{\mathrm{gate}} > 0, \qquad \forall t \in [T_1, t_\star), \tag{204}$$

for some constant $c_{\mathrm{gate}} \in (0, 1]$.

The effective coupling coefficient is thus

$$\hat{a}_{2,t}^{\mathrm{eff}} = \Theta\Big(\frac{\alpha_2^2\eta}{d}\Big) \cdot \hat{p}_{i,j}^{(t)} \ge \Theta\Big(\frac{\alpha_2^2\eta}{d}\Big). \tag{205}$$

Iterating the symmetric 2D system yields

$$\left|\frac{\langle \hat{w}_i^{(t)}, \mathbf{M}_{2j}\rangle + \langle \hat{v}_i^{(t)}, \mathbf{M}_{2j}\rangle}{2}\right| \ge \Big(1 + \Theta\Big(\frac{\alpha_2^2\eta}{d}\Big)\Big)^{t-T_1} \cdot \Theta\Big(\frac{\sigma_0}{\sqrt{d}}\Big) - \sum_{s=T_1}^{t-1} |\widehat{\mathrm{Err}}_s|. \tag{206}$$

G.5. Hitting Time and Definition of $S_{2j,\mathrm{sure}}$ (Joint Scheduling)

We define the sure set $S_{2j,\mathrm{sure}}$ under the joint scheduling so that for each $i \in S_{2j,\mathrm{sure}}$, the coordinate j is guaranteed to be the first $\mathbf{M}_2$-direction that crosses the bias. Additionally, we require $i \notin S_1$, ensuring that neuron i is frozen during Stage I.

Using (206), the hitting condition (196) is satisfied when

$$\Big(1 + \Theta\Big(\frac{\alpha_2^2\eta}{d}\Big)\Big)^{t_\star - T_1} \cdot \Theta\Big(\frac{\sigma_0}{\sqrt{d}}\Big) \ge \hat{b}_i^{(t_\star)}. \tag{207}$$

Since $\hat{b}_i^{(t_\star)} = (1 + \eta/d)^{t_\star - T_1}\hat{b}_i^{(T_1)}$, the condition holds when Stage II is sufficiently long. With the choice

$$T_2 = \Theta\Big(\frac{d^{1+2c_0}\log d}{\eta}\Big), \tag{208}$$

we have

$$\Big(1 + \Theta\Big(\frac{\alpha_2^2\eta}{d}\Big)\Big)^{T_2} = \mathrm{poly}(d), \tag{209}$$

and the sure-set condition selects those neurons for which the net ratio in (207) crosses 1.

G.6. Regime B: Exponential Growth After Crossing the Bias

Fix a neuron $i \in S_{2j,\mathrm{sure}}$, and let $t_\star \in [T_1, T_1 + T_2]$ denote the first time such that the alignment along $\mathbf{M}_{2j}$ reaches the activation scale.

After crossing the threshold, the $\mathbf{M}_{2j}$ component becomes the dominant contributor to the pre-activation. By Lemma 4, for all $t \geq t_\star$,

$$\mathbf{1}\{\hat{w}_i^{(t)\top} x_\tau \geq \hat{b}_i^{(t)}\} = 1 \quad \text{whenever } z_{2j} = 1, \tag{210}$$

with probability at least $1 - e^{-\Omega(d_1)}$.

The dynamics along $\mathbf{M}_{2j}$ reduce to the symmetric linear system

$$\begin{bmatrix} \langle \hat{w}_i^{(t+1)}, \mathbf{M}_{2j} \rangle \\ \langle \hat{v}_i^{(t+1)}, \mathbf{M}_{2j} \rangle \end{bmatrix} = \begin{bmatrix} 1 & \hat{a} \\ \hat{a} & 1 \end{bmatrix} \begin{bmatrix} \langle \hat{w}_i^{(t)}, \mathbf{M}_{2j} \rangle \\ \langle \hat{v}_i^{(t)}, \mathbf{M}_{2j} \rangle \end{bmatrix} \pm \widehat{\mathrm{Err}}_t, \qquad \hat{a} = \Theta\left(\frac{\alpha_2^2 \eta}{d}\right). \tag{211}$$

Diagonalizing the update matrix, we obtain for all $t \geq t_\star$,

$$|\langle \hat{w}_i^{(t)}, \mathbf{M}_{2j} \rangle| \;\geq\; \left(1 + \Theta\left(\frac{\alpha_2^2 \eta}{d}\right)\right)^{t - t_\star} \cdot |\langle \hat{w}_i^{(t_\star)}, \mathbf{M}_{2j} \rangle| - o\Big(\|\hat{w}_i^{(t)}\|_2\Big). \tag{212}$$

G.7. Purity of $\mathbf{M}_2$ Alignment (Key Difference from Denoise-Only)

The crucial difference from the denoise-only case is that the $\mathbf{M}_1$ components remain negligible throughout training.

Since $i \in S_{2j,\mathrm{sure}}$ was frozen during Stage I, we have

$$|\langle \hat{w}_i^{(T_1)}, \mathbf{M}_{1j'} \rangle| = O\left(\frac{\sigma_0}{\sqrt{d}}\right), \qquad \forall j' \in [d]. \tag{213}$$

During Stage II, the $\mathbf{M}_1$ components receive negligible updates because the predictor $g^{(t)}$ has already learned the $\mathbf{M}_1$ structure accurately. By the same argument as in Section F.6,

$$\left|\langle \hat{w}_i^{(t+1)}, \mathbf{M}_{1j'} \rangle - \langle \hat{w}_i^{(t)}, \mathbf{M}_{1j'} \rangle\right| \leq \frac{1}{\mathrm{poly}(d)} \|\hat{w}_i^{(t)}\|_2, \qquad \forall t \in [T_1, T_1 + T_2]. \tag{214}$$

Summing over t and using $T_2 = \mathrm{poly}(d)$, the total drift of $\mathbf{M}_1$ components during Stage II is at most $O(\sigma_0)$, which remains negligible compared to the final norm $\|\hat{w}_i^{(T_1+T_2)}\|_2 = \Theta(1)$.

Therefore, at the end of Stage II,

$$|\langle \hat{w}_i^{(T_1+T_2)}, \mathbf{M}_{1j'} \rangle|^2 = o(1)\, \|\hat{w}_i^{(T_1+T_2)}\|_2^2, \qquad \forall j' \in [d]. \tag{215}$$

Since the $\mathbf{M}_{2j}$ component grows exponentially while all other components ($\mathbf{M}_1$, other $\mathbf{M}_2$ directions, and orthogonal directions) remain negligible, the final norm is dominated by the $\mathbf{M}_{2j}$ alignment:

$$|\langle \hat{w}_i^{(T_1+T_2)}, \mathbf{M}_{2j} \rangle|^2 = (1 - o(1))\, \|\hat{w}_i^{(T_1+T_2)}\|_2^2. \tag{216}$$

This establishes the **pure** $\mathbf{M}_{2j}$ alignment claimed in Theorem 2.

G.8. Summary of Stage II (Joint Scheduling)

At the end of Stage II ($t = T_1 + T_2$), with $T_2 = \Theta\Big(\frac{d^{1+2c_0} \log d}{\eta}\Big)$, the alignment properties under the joint scheduling are summarized as follows.

For $i \in S_{1j,\mathrm{sure}}$: the $\mathbf{M}_1$ alignment learned in Stage I remains stable.

$$|\langle w_i^{(T_1+T_2)}, \mathbf{M}_{1j} \rangle|^2 \geq (1 - o(1))\, \|w_i^{(T_1+T_2)}\|_2^2, \tag{217}$$

and all other $\mathbf{M}_1$ and $\mathbf{M}_2$ coordinates remain negligible. These neurons achieve pure $\mathbf{M}_{1j}$ representations.

For $i \in S_{2j,\text{sure}}$: the neuron was frozen during Stage I and enters the exponential-growth regime along $\mathbf{M}_{2j}$ during Stage II. Since the $\mathbf{M}_1$ components remain at initialization scale,

$$|\langle \hat{w}_i^{(T_1+T_2)}, \mathbf{M}_{2j} \rangle|^2 = (1 - o(1)) \, \|\hat{w}_i^{(T_1+T_2)}\|_2^2, \tag{218}$$

$$|\langle \hat{w}_i^{(T_1+T_2)}, \mathbf{M}_{1j'} \rangle|^2 = o(1) \, \|\hat{w}_i^{(T_1+T_2)}\|_2^2, \qquad \forall j' \in [d]. \tag{219}$$

These neurons achieve **pure** $\mathbf{M}_{2j}$ representations.

For orthogonal directions: for all neurons $i \in [m]$,

$$\|(I - \mathbf{M}_1\mathbf{M}_1^\top - \mathbf{M}_2\mathbf{M}_2^\top)\hat{w}_i^{(T_1+T_2)}\|_2^2 \le O\left(\frac{1}{d_1}\right) \|\hat{w}_i^{(T_0)}\|_2^2. \tag{220}$$

In contrast to the denoise-only protocol (Section F), the joint scheduling produces **pure** $\mathbf{M}_2$ neurons. The sparsity schedule (freezing $\mathbf{M}_2$-targeted neurons during Stage I) prevents the accumulation of $\mathbf{M}_1$ alignment, enabling these neurons to specialize exclusively to $\mathbf{M}_2$ directions. Combined with the pure $\mathbf{M}_1$-aligned neurons from Stage I, the final network contains specialized neurons for both $\mathbf{M}_1$ and $\mathbf{M}_2$ directions, each achieving near-perfect single-direction alignment.

H. M_2 Learning under Standard Training (Entangled Representations)

H.1. Setup and Goal

Under **standard training** (without any high-to-low scheduling): the noise level is sampled uniformly from the full range throughout training,

$$\tau \sim \text{Unif}[\tau_{\min}, \tau_{\max}],$$

where

$$\tau_{\min} = \Theta(1/d_1), \quad \tau_{\max} = \Theta(\alpha_1/\sqrt{\log d}).$$

Crucially, **all neurons participate in training from the beginning**, and the noise level can be as large as $\tau_{\max} > \alpha_2$. This means that when $\tau > \alpha_2$, the $\mathbf{M}_2$ signal is overwhelmed by noise, preventing reliable feature discrimination.

In this section, we prove Theorem 4: under standard training, neurons that learn $\mathbf{M}_2$ features inevitably encode **multiple directions simultaneously**, leading to entangled representations. The key mechanism is that high-noise samples trigger the **Mixed Regime** of Lemma 4, which destroys gating separation and causes multiple $\mathbf{M}_2$ directions to receive gradient updates concurrently.

H.2. Projected Dynamics Along $\mathbf{M}_{2j}$

We use the same projection framework as in Section F. For any direction $\mathbf{M}_{2j}$, the SGD update projected onto $\mathbf{M}_{2j}$ satisfies:

$$\begin{aligned}
\langle \bar{w}_i^{(t+1)}, \mathbf{M}_{2j} \rangle &= \langle \bar{w}_i^{(t)}, \mathbf{M}_{2j} \rangle + \eta \, \langle \bar{v}_i^{(t)}, \mathbf{M}_{2j} \rangle \, \alpha_2^2 z_{2j}^2 \, \mathbf{1}\{\bar{w}_i^{(t)\top} x_\tau \ge b_i^{(t)}\} \pm \text{Err}_t, \\
\langle \bar{v}_i^{(t+1)}, \mathbf{M}_{2j} \rangle &= \langle \bar{v}_i^{(t)}, \mathbf{M}_{2j} \rangle + \eta \, \langle \bar{w}_i^{(t)}, \mathbf{M}_{2j} \rangle \, \alpha_2^2 z_{2j}^2 \, \mathbf{1}\{\bar{w}_i^{(t)\top} x_\tau \ge b_i^{(t)}\} \pm \text{Err}_t.
\end{aligned} \tag{221}$$

Define the gate frequency conditioned on $z_{2j} = 1$:

$$\bar{p}_{i,j}^{(t)} = \Pr\left(\bar{w}_i^{(t)\top} x_\tau \ge b_i^{(t)} \,\middle|\, z_{2j} = 1\right). \tag{222}$$

The effective coupling coefficient for direction $\mathbf{M}_{2j}$ is

$$\bar{a}_{j,t}^{\text{eff}} = \Theta\left(\frac{\alpha_2^2 \eta}{d}\right) \cdot \bar{p}_{i,j}^{(t)}. \tag{223}$$

H.3. High-Noise Regime: Failure of Gating Separation

The critical difference from the high-to-low scheduling is that standard training samples noise from the full range $[\tau_{\min}, \tau_{\max}]$, including high-noise samples where $\tau > \alpha_2/\sqrt{\log d}$.

Consider the pre-activation decomposition when $z_{2j} = 1$:

$$\bar{w}_i^{(t)\top} x_\tau = \underbrace{\alpha_2 \langle \bar{w}_i^{(t)}, \mathbf{M}_{2j} \rangle}_{\text{signal from } \mathbf{M}_{2j}} + \underbrace{\sum_{j' \neq j} \alpha_2 z_{2j'} \langle \bar{w}_i^{(t)}, \mathbf{M}_{2j'} \rangle}_{\text{other } \mathbf{M}_2 \text{ directions}} + \underbrace{\tau \langle \bar{w}_i^{(t)}, \epsilon \rangle}_{\text{noise}}. \tag{224}$$

For high-noise samples with $\tau = \Theta(\tau_{\max}) = \Theta(\alpha_1/\sqrt{\log d}) \gg \alpha_2$, the noise term dominates the $\mathbf{M}_2$ signal contributions. Specifically, the signal-to-noise ratio for $\mathbf{M}_2$ features satisfies:

$$\frac{\alpha_2 |\langle \bar{w}_i^{(t)}, \mathbf{M}_{2j} \rangle|}{\tau \|\bar{w}_i^{(t)}\|_2 / \sqrt{d_1}} = O\left(\frac{\alpha_2 \sqrt{d_1}}{\tau_{\max}}\right) = O\left(\frac{\alpha_2 \sqrt{d_1 \log d}}{\alpha_1}\right) = o(1), \tag{225}$$

since $\alpha_2 = \Theta(1/d^{c_0})$ and $\alpha_1 = \Theta(1)$.

By Lemma 4 (Regime (B), Mixed Regime), when the noise level satisfies $\sqrt{\tau} \geq \max\{|\Delta_1|, |\Delta_2|\}/(\sqrt{2\pi}\delta)$, the activation probabilities are close to random guessing:

$$\left|\bar{p}_{i,j}^{(t)} - \tfrac{1}{2}\right| \leq o(1), \qquad \left|\bar{p}_{i,j'}^{(t)} - \tfrac{1}{2}\right| \leq o(1), \qquad \forall j, j' \in [d]. \tag{226}$$

This means that for high-noise samples, **all** $\mathbf{M}_2$ directions have approximately equal probability of triggering the gate. Unlike the high-to-low setting where gating separation ensures only the target direction $\mathbf{M}_{2j}$ activates (cf. Eqs. (182)–(183)), standard training lacks this separation mechanism.

H.4. Multi-Direction Gradient Accumulation

Since gating separation fails for high-noise samples, multiple $\mathbf{M}_2$ directions receive gradient updates simultaneously. Define the set of directions that neuron i learns:

$$\mathcal{N}_i = \left\{ j \in [d] : |\langle \bar{w}_i^{(t)}, \mathbf{M}_{2j} \rangle| \geq \Omega\left(\frac{\|\bar{w}_i^{(t)}\|_2}{\sqrt{d}}\right) \text{ for } t \geq T_1 \right\}. \tag{227}$$

We now show that $|\mathcal{N}_i| = \Theta(\sqrt{d})$ for neurons that learn $\mathbf{M}_2$ features.

Consider the gradient contribution from direction $\mathbf{M}_{2j}$ when $z_{2j} = 1$. For high-noise samples (which constitute a constant fraction of the training distribution), the effective coupling is:

$$\bar{a}_{j,t}^{\text{eff}} = \Theta\left(\frac{\alpha_2^2 \eta}{d}\right) \cdot \bar{p}_{i,j}^{(t)} = \Theta\left(\frac{\alpha_2^2 \eta}{d}\right) \cdot \left(\frac{1}{2} \pm o(1)\right). \tag{228}$$

Since all directions $j \in [d]$ have the same effective coupling rate $\Theta(\alpha_2^2 \eta/d)$ under high-noise samples, directions with similar initial alignments grow at comparable rates. Crucially, this means the relative magnitudes of alignments are approximately preserved during training.

At initialization, for any neuron i, the projections onto $\mathbf{M}_2$ directions are i.i.d. Gaussian:

$$\langle \bar{w}_i^{(0)}, \mathbf{M}_{2j} \rangle \sim \mathcal{N}(0, \sigma_0^2), \qquad \forall j \in [d]. \tag{229}$$

We use Gaussian order statistics to determine how many directions have comparable initial alignment. Let $a_j = |\langle \bar{w}_i^{(0)}, \mathbf{M}_{2j} \rangle|$ denote the initial alignment magnitude. By standard extreme value theory, the maximum satisfies

$$a_{\max} = \max_{j \in [d]} a_j = \Theta(\sigma_0 \sqrt{\log d}) \tag{230}$$

with high probability.

We count the number of directions whose initial alignment is within a constant factor of the maximum. By the Gaussian tail bound, for threshold $t = \sigma_0\sqrt{2\log d}/C$ with constant $C > 1$:

$$\Pr(|X| \geq t) = 2(1 - \Phi(t/\sigma_0)) \approx \sqrt{\frac{2}{\pi}} \cdot \frac{e^{-t^2/(2\sigma_0^2)}}{t/\sigma_0} = \Theta\left(\frac{1}{\sqrt{\log d}}\right) \cdot d^{-1/C^2}. \tag{231}$$

The expected number of directions exceeding this threshold is:

$$\mathbb{E}\left[\left|\left\{j : a_j \geq \frac{a_{\max}}{C}\right\}\right|\right] = d \cdot \Pr(|X| \geq t) = \Theta\left(\frac{d^{1-1/C^2}}{\sqrt{\log d}}\right). \tag{232}$$

Taking $C = \sqrt{2}$ (directions within factor $1/\sqrt{2}$ of the maximum), we have $1 - 1/C^2 = 1/2$, yielding:

$$|\mathcal{N}_i| = \Theta\left(\sqrt{d/\log d}\right). \tag{233}$$

Since all directions in $\mathcal{N}_i$ have initial alignments within a constant factor of each other, and they grow at the same rate under high-noise samples, they remain comparable throughout training. This establishes that $\Theta(\sqrt{d})$ directions are learned simultaneously.

H.5. Entangled Weight Decomposition

Iterating the dynamics in (221) with the multi-direction coupling (228), we obtain for $t \geq T_1$:

$$|\langle \bar{w}_i^{(t)}, \mathbf{M}_{2j}\rangle| \geq \left(1 + \Theta\left(\frac{\alpha_2^2\eta}{d}\right)\right)^{t-T_1} \cdot |\langle \bar{w}_i^{(T_1)}, \mathbf{M}_{2j}\rangle| - o(\|\bar{w}_i^{(t)}\|_2), \qquad \forall j \in \mathcal{N}_i. \tag{234}$$

Since all directions in $\mathcal{N}_i$ grow at comparable rates, the final weights decompose as a mixture:

$$\bar{w}_i^{(t)} = \sum_{j\in\mathcal{N}_i} \bar{\beta}_{i,j}\mathbf{M}_{2j} + \bar{\mathbf{r}}_i, \qquad \bar{v}_i^{(t)} = \sum_{j\in\mathcal{N}_i} \bar{\beta}_{i,j}\mathbf{M}_{2j} + \bar{\mathbf{s}}_i, \tag{235}$$

where the coefficients satisfy:

$$\bar{\beta}_{i,j}^2 = \Theta\left(\frac{\|\bar{w}_i^{(t)}\|_2^2 + \|\bar{v}_i^{(t)}\|_2^2}{|\mathcal{N}_i|}\right), \qquad \forall j \in \mathcal{N}_i. \tag{236}$$

H.6. Residual Error Analysis

The residual $\bar{\mathbf{r}}_i$ consists of: (i) components along $\mathbf{M}_1$ directions accumulated during training, (ii) components along $\mathbf{M}_2$ directions outside $\mathcal{N}_i$, (iii) orthogonal noise components.

Unlike the joint scheduling case where $\mathbf{M}_2$-targeted neurons are frozen during Stage I (yielding $o(1)$ residual), standard training allows all neurons to accumulate $\mathbf{M}_1$ alignment throughout. Moreover, the lack of gating separation means that even directions outside $\mathcal{N}_i$ receive non-negligible gradient contributions.

Combining these effects, the residual satisfies:

$$\|\bar{\mathbf{r}}_i\|_2^2, \|\bar{\mathbf{s}}_i\|_2^2 = \Theta\left(\|\bar{w}_i^{(t)}\|_2^2 + \|\bar{v}_i^{(t)}\|_2^2\right). \tag{237}$$

For neurons $i' \notin \bigcup_j \mathcal{N}_{i'}$ (those that do not specialize in any $\mathbf{M}_2$ direction), all $\mathbf{M}_2$ projections remain small:

$$|\langle \bar{w}_{i'}^{(t)}, \mathbf{M}_{2j}\rangle| \leq O\left(\frac{\|\bar{w}_{i'}^{(t)}\|_2^2 + \|\bar{v}_{i'}^{(t)}\|_2^2}{d}\right), \qquad \forall j \in [d]. \tag{238}$$

H.7. Summary: Entanglement under Standard Training

The analysis above establishes the key claims of Theorem 4:

(1) Multi-direction alignment. For neurons that learn $\mathbf{M}_2$ features, the weights align to multiple directions simultaneously, with $|\mathcal{N}_i| = \Theta(\sqrt{d})$. This is a direct consequence of the Mixed Regime (Lemma 4(B)) destroying gating separation under high-noise samples, combined with Gaussian order statistics showing that $\Theta(\sqrt{d})$ directions have comparable initial alignment.

(2) Coefficient scaling. Each direction in $\mathcal{N}_i$ contributes comparably to the total norm, with $\bar{\beta}_{i,j}^2 = \Theta((\|\bar{w}_i^{(t)}\|_2^2 + \|\bar{v}_i^{(t)}\|_2^2)/|\mathcal{N}_i|)$.

(3) Significant residual. The residual error is of the same order as the total norm, $\|\bar{\mathbf{r}}_i\|_2^2 = \Theta(\|\bar{w}_i^{(t)}\|_2^2)$, reflecting the accumulated interference from $\mathbf{M}_1$ alignment and cross-direction contamination.

Comparison with the high-to-low scheduling. The fundamental difference lies in the gating behavior:

	High-to-Low (Stage II)	Standard Training
Noise range	$[\tau_{\min}, \tau_1]$ with $\tau_1 < \alpha_2$	$[\tau_{\min}, \tau_{\max}]$ with $\tau_{\max} > \alpha_2$
Lemma regime	Separable (A)	Mixed (B)
Gating separation	Yes	No
$\mathbf{M}_2$ alignment	Pure (single direction)	Entangled (multiple directions)

This completes the proof of Theorem 4.

I. Convergence of the Diffusion Training Dynamics

In this section, we provide the convergence analysis for the diffusion model training. Building upon the alignment results from Stage I and Stage II, we show that the network parameters converge to a stationary point that corresponds to the optimal reconstruction of the clean signal structure.

I.1. Indicator Stability and Near-Perfect Alignment

We first establish that in the final training phase, the activation patterns of the neurons become stable. Fix a feature index $j \in [d]$.

For any neuron $i \in S_{1j,\text{sure}}$, the Stage I analysis implies that the weights aligned with $\mathbf{M}_{1j}$ satisfy:

$$\frac{\langle w_i^{(T_1)}, \mathbf{M}_{1j}\rangle}{\|w_i^{(T_1)}\|_2} = 1 - o(1), \qquad \frac{\langle v_i^{(T_1)}, \mathbf{M}_{1j}\rangle}{\|v_i^{(T_1)}\|_2} = 1 - o(1). \tag{239}$$

Moreover, for all $k \neq j$, the off-support projections are negligible:

$$\frac{|\langle w_i^{(T_1)}, \mathbf{M}_{1k}\rangle|}{\|w_i^{(T_1)}\|_2} \leq o(1), \qquad \frac{|\langle v_i^{(T_1)}, \mathbf{M}_{1k}\rangle|}{\|v_i^{(T_1)}\|_2} \leq o(1). \tag{240}$$

Recall that the input is given by $x_\tau = x_0 + \tau\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. Standard concentration inequalities (Lemma 1) imply that $|\langle w_i, \epsilon\rangle|$ is bounded by $O(\|w_i\|_2/\sqrt{d})$ with high probability. Since the bias $b_i^{(t)}$ grows sublinearly compared to the signal margin $\langle w_i^{(t)}, x_\tau\rangle$, the gating condition satisfies:

$$\Pr\left(\mathbf{1}\{w_i^{(t)\top} x_\tau \geq b_i^{(t)}\} = 1\right) \geq 1 - e^{-\Omega(d)}, \qquad \forall t \in [T_1, T_1 + T_2], \tag{241}$$

whenever the sample contains the feature $z_{1j} = 1$.

Similarly, for any neuron $i \in S_{2j,\mathrm{sure}}$, the Stage II dynamics ensure that the dominant component is $\mathbf{M}_{2j}$. Specifically:

$$\frac{\langle w_i^{(T_1+T_2)}, \mathbf{M}_{2j}\rangle}{\|w_i^{(T_1+T_2)}\|_2} = 1 - o(1), \qquad \frac{\langle v_i^{(T_1+T_2)}, \mathbf{M}_{2j}\rangle}{\|v_i^{(T_1+T_2)}\|_2} = 1 - o(1). \tag{242}$$

All off-support coefficients remain at most $O(1/\sqrt{d})$. Consequently, when the data sample contains the feature $z_{2j} = 1$, it holds with high probability that:

$$\mathbf{1}\{w_i^{(t)\top} x_\tau \geq b_i^{(t)}\} = 1, \qquad \forall t \geq T_1 + T_2. \tag{243}$$

Equations (241) and (243) show that for sure-set neurons, the ReLU activation state is frozen with high probability. Under this condition, the mapping $(w_i, v_i) \mapsto g(x_\tau)$ behaves linearly, and the diffusion loss reduces to a locally quadratic function of (w_i, v_i), ensuring convexity and smoothness.

I.2. Identification of the Optimal Aligned Solution

We analyze the convergence targets for Stage I and Stage II separately, reflecting the sequential learning nature of the diffusion process.

Stage I: Learning $\mathbf{M}_1$. Since $\alpha_1 = \Theta(1)$ and $\alpha_2 = \Theta(1/d^{c_0})$, the effective target in Stage I is dominated by $\alpha_1 z_{1j}\mathbf{M}_{1j}$. Under the stable indicator (241), the restricted loss for neuron $i \in S_{1j,\mathrm{sure}}$ is:

$$L_{\mathrm{DM},1j}(w_i, v_i) = \mathbb{E}\Big[\tfrac{1}{2}\|v_i^\top (w_i^\top x_\tau - b_i) - \alpha_1 z_{1j}\mathbf{M}_{1j}\|_2^2\Big]. \tag{244}$$

Because x_τ lies in the subspace spanned by $\mathbf{M}_1$ up to noise terms, the minimizer is the rank-one solution aligned with the dictionary atom:

$$w_i^* = \beta_{i,1j}^* \mathbf{M}_{1j}, \qquad v_i^* = \gamma_{i,1j}^* \mathbf{M}_{1j}, \tag{245}$$

where $\beta_{i,1j}^*$ and $\gamma_{i,1j}^*$ are scalar coefficients determined by the signal-to-noise ratio.

Stage II: Learning $\mathbf{M}_2$. After Stage I, the $\mathbf{M}_1$ component has been fitted. The residual signal corresponds to the component $\alpha_2 z_{2j}\mathbf{M}_{2j}$. For neurons $i \in S_{2j,\mathrm{sure}}$, the restricted quadratic loss becomes:

$$L_{\mathrm{DM},2j}(w_i, v_i) = \mathbb{E}\Big[\tfrac{1}{2}\|v_i^\top (w_i^\top x_\tau - b_i) - \alpha_2 z_{2j}\mathbf{M}_{2j}\|_2^2\Big]. \tag{246}$$

Due to indicator stability (243), this loss is quadratic. Since $\mathbf{M}_{2j}$ dominates the residual, the unique minimizer aligns purely with $\mathbf{M}_{2j}$:

$$w_i^* = \beta_{i,2j}^* \mathbf{M}_{2j}, \qquad v_i^* = \gamma_{i,2j}^* \mathbf{M}_{2j}. \tag{247}$$

I.3. Smoothness and Gradient Descent Convergence

For all sure-set neurons, the restricted losses $L_{\mathrm{DM},1j}$ and $L_{\mathrm{DM},2j}$ are quadratic and convex. To apply standard optimization bounds, we first verify the smoothness condition.

The Hessian of the loss is dominated by the covariance of the input x_τ. Using the bounds $\mathbb{E}[\|x_\tau\|_2^2] = \Theta(1)$ and $\|w_i^{(t)}\|_2 = \|v_i^{(t)}\|_2 = \Theta(1)$, the smoothness constant ℓ_i satisfies:

$$\ell_i = O(\|v_i^{(t)}\|_2^2 \mathbb{E}[\|x_\tau\|_2^2] + \|w_i^{(t)}\|_2^2 \mathbb{E}[\|x_\tau\|_2^2]) = \Theta(1). \tag{248}$$

Let $L = \max_i \ell_i = \Theta(1)$. We consider the gradient descent update rules with learning rate $\eta \leq 1/L$:

$$w_i^{(t+1)} = w_i^{(t)} - \eta \nabla_{w_i} L_{\mathrm{DM},j}^{(t)}, \qquad v_i^{(t+1)} = v_i^{(t)} - \eta \nabla_{v_i} L_{\mathrm{DM},j}^{(t)}. \tag{249}$$

By the Descent Lemma for L-smooth functions, we have:

$$L_{\mathrm{DM},j}(w_i^{(t+1)}, v_i^{(t+1)}) \leq L_{\mathrm{DM},j}(w_i^{(t)}, v_i^{(t)}) - \tfrac{\eta}{2}\|\nabla L_{\mathrm{DM},j}(w_i^{(t)}, v_i^{(t)})\|_2^2. \tag{250}$$

Since the problem is locally strongly convex with parameter $\mu = \Theta(1)$, it satisfies the Polyak-Łojasiewicz (PL) condition:

$$\|\nabla L_{\mathrm{DM},j}(w_i^{(t)}, v_i^{(t)})\|_2^2 \geq 2\mu\big(L_{\mathrm{DM},j}(w_i^{(t)}, v_i^{(t)}) - L_{\mathrm{DM},j}(w_i^*, v_i^*)\big). \tag{251}$$

Averaging (250) over $t = T_1 + T_2, \ldots, T_1 + T_2 + T_{\mathrm{DM}} - 1$ and substituting (251) gives the telescope bound:

$$\frac{1}{T_{\mathrm{DM}}} \sum_{t=T_1+T_2}^{T_1+T_2+T_{\mathrm{DM}}-1} L_{\mathrm{DM},j}(w_i^{(t)}, v_i^{(t)}) \le L_{\mathrm{DM},j}(w_i^*, v_i^*) + \frac{\Delta_{T_1+T_2}}{\mu \cdot T_{\mathrm{DM}}}, \tag{252}$$

where

$$\Delta_{T_1+T_2} = L_{\mathrm{DM},j}(w_i^{(T_1+T_2)}, v_i^{(T_1+T_2)}) - L_{\mathrm{DM},j}(w_i^*, v_i^*) = \Theta(1). \tag{253}$$

Using $T_{\mathrm{DM}} = \Theta(d^2)$ and $\mu = \Theta(1)$, we obtain:

$$\frac{\Delta_{T_1+T_2}}{\mu \cdot T_{\mathrm{DM}}} = \Theta\Big(\frac{1}{d^2}\Big). \tag{254}$$

I.4. Noise-Regime-Dependent Recovery

The final loss value depends critically on the noise level τ. We analyze the three regimes separately.

Low-noise regime ($\tau \ll \alpha_2/\sqrt{\log d}$). In this regime, both $\mathbf{M}_1$ and $\mathbf{M}_2$ can be reliably recovered. The loss satisfies:

$$L_{\mathrm{DM}} = \mathbb{E}_{x_0,\tau} \tfrac{1}{2}\|g(x_\tau) - x_0\|_2^2 \le \Theta\Big(\frac{1}{d^2}\Big). \tag{255}$$

Intermediate-noise regime ($\alpha_2/\sqrt{\log d} \ll \tau \ll \alpha_1/\sqrt{\log d}$). In this regime, only $\mathbf{M}_1$ can be reliably recovered, while $\mathbf{M}_2$ remains corrupted by noise. The optimal denoiser recovers $\mathbf{M}_1$ but returns zero for $\mathbf{M}_2$ directions. The residual loss is dominated by the unrecovered $\mathbf{M}_2$ component:

$$L_{\mathrm{DM}} = \Theta(\alpha_2^2 d) = \Theta\Big(\frac{1}{d^{2c_0}}\Big). \tag{256}$$

High-noise regime ($\tau \gg \alpha_1/\sqrt{\log d}$). In this regime, the noise dominates both signal components. The optimal denoiser is approximately zero, yielding:

$$L_{\mathrm{DM}} = \Theta(1). \tag{257}$$

I.5. Summary

At the end of training ($t = T_1 + T_2 + T_{\mathrm{DM}}$), the alignment and loss properties are as follows:

- For $i \in S_{1j,\mathrm{sure}}$: the neuron converges to the optimal solution (245) aligned with $\mathbf{M}_{1j}$.
- For $i \in S_{2j,\mathrm{sure}}$: the neuron converges to the optimal solution (247) aligned with $\mathbf{M}_{2j}$.
- The expected diffusion loss depends on the noise regime:

$$L_{\mathrm{DM}} = \begin{cases} \Theta(1/d^2) & \text{if } \tau \ll \alpha_2/\sqrt{\log d}, \\ \Theta(1/d^{2c_0}) & \text{if } \alpha_2/\sqrt{\log d} \ll \tau \ll \alpha_1/\sqrt{\log d}, \\ \Theta(1) & \text{if } \tau \gg \alpha_1/\sqrt{\log d}. \end{cases} \tag{258}$$

This confirms that the diffusion model learns the hierarchical structure of the data: (1) $\mathbf{M}_1$ is recovered at high and intermediate noise levels. (2) $\mathbf{M}_2$ is only recovered at low noise levels, consistent with its weaker signal strength $\alpha_2 = \Theta(1/d^{c_0})$.

J. Generalization

This section provides the proof of Theorem 5. We analyze the generalization error of the learned denoisers g_{joint} (trained with joint denoising-sparsity scheduling) and g_{std} (trained with standard training) across different noise regimes.

J.1. Score Function and Optimal Denoiser

We begin by formalizing the forward corruption process used in diffusion training. The noisy observation is generated as

$$x_\tau = x_0 + \tau\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{259}$$

so that, conditioned on the clean signal x_0, the likelihood of the noisy sample x_τ is Gaussian with covariance $\tau^2 I$:

$$p(x_\tau \mid x_0) = \frac{1}{(2\pi\tau^2)^{d_1/2}} \exp\left(-\frac{\|x_\tau - x_0\|_2^2}{2\tau^2}\right). \tag{260}$$

Let $p_\tau(x_\tau)$ denote the marginal density of the noisy sample, integrated over the clean-data distribution:

$$p_\tau(x_\tau) = \int p(x_\tau \mid x_0)\, p(x_0)\, dx_0. \tag{261}$$

The target object of interest in reverse-time diffusion is the score function

$$s_\tau^*(x_\tau) = \nabla_{x_\tau} \log p_\tau(x_\tau), \tag{262}$$

which appears explicitly in the reverse-time SDE that governs generative sampling.

To derive a closed-form expression for the score under the additive Gaussian model, we differentiate the marginal density:

$$\nabla_{x_\tau} p_\tau(x_\tau) = \nabla_{x_\tau} \int p(x_\tau \mid x_0)\, p(x_0)\, dx_0 = \int \nabla_{x_\tau} p(x_\tau \mid x_0)\, p(x_0)\, dx_0. \tag{263}$$

The Gaussian likelihood satisfies

$$\nabla_{x_\tau} p(x_\tau \mid x_0) = p(x_\tau \mid x_0) \left(-\frac{x_\tau - x_0}{\tau^2}\right). \tag{264}$$

Substituting this expression yields

$$\begin{aligned} \nabla_{x_\tau} p_\tau(x_\tau) &= \int p(x_\tau \mid x_0)\, p(x_0)\, \frac{x_0 - x_\tau}{\tau^2}\, dx_0 \\ &= \frac{1}{\tau^2} \left(\int x_0\, p(x_\tau \mid x_0)\, p(x_0)\, dx_0 - x_\tau \int p(x_\tau \mid x_0)\, p(x_0)\, dx_0 \right). \end{aligned} \tag{265}$$

The second integral is simply $p_\tau(x_\tau)$. Using Bayes' rule for the first integral,

$$p(x_0 \mid x_\tau) = \frac{p(x_\tau \mid x_0)\, p(x_0)}{p_\tau(x_\tau)}, \tag{266}$$

we obtain

$$\int x_0\, p(x_\tau \mid x_0)\, p(x_0)\, dx_0 = p_\tau(x_\tau)\, \mathbb{E}[x_0 \mid x_\tau]. \tag{267}$$

Therefore,

$$\nabla_{x_\tau} p_\tau(x_\tau) = \frac{p_\tau(x_\tau)}{\tau^2} \big(\mathbb{E}[x_0 \mid x_\tau] - x_\tau\big), \tag{268}$$

and dividing by $p_\tau(x_\tau)$ gives Tweedie's formula:

$$s_\tau^*(x_\tau) = \nabla_{x_\tau} \log p_\tau(x_\tau) = \frac{\mathbb{E}[x_0 \mid x_\tau] - x_\tau}{\tau^2}. \tag{269}$$

The diffusion-training objective is typically formulated as the mean squared error (MSE) regression of the clean signal x_0 on the noisy observation x_τ:

$$L_{\mathrm{DM}}(g) = \mathbb{E}_{x_0, x_\tau} \left[\frac{1}{2} \|g(x_\tau) - x_0\|_2^2\right]. \tag{270}$$

To characterize the population minimizer of this objective, we employ the bias-variance decomposition. We add and subtract the conditional expectation $\mathbb{E}[x_0 \mid x_\tau]$ inside the norm:

$$\begin{aligned} 2L_{\mathrm{DM}}(g) &= \mathbb{E}\left[\|g(x_\tau) - \mathbb{E}[x_0 \mid x_\tau] + \mathbb{E}[x_0 \mid x_\tau] - x_0\|_2^2\right] \\ &= \underbrace{\mathbb{E}\left[\|g(x_\tau) - \mathbb{E}[x_0 \mid x_\tau]\|_2^2\right]}_{\text{Estimation Error}} + \underbrace{\mathbb{E}\left[\|\mathbb{E}[x_0 \mid x_\tau] - x_0\|_2^2\right]}_{\text{Irreducible Error}} \\ &\quad + \underbrace{2\mathbb{E}\left[\langle g(x_\tau) - \mathbb{E}[x_0 \mid x_\tau], \mathbb{E}[x_0 \mid x_\tau] - x_0\rangle\right]}_{\text{Cross Term}}. \end{aligned} \tag{271}$$

The cross term vanishes due to the orthogonality property of the conditional expectation. Specifically, by applying the law of iterated expectations and conditioning on x_τ, we have:

$$\begin{aligned} \mathbb{E}_{x_0,x_\tau}[\dots] &= \mathbb{E}_{x_\tau}\left[\langle g(x_\tau) - \mathbb{E}[x_0 \mid x_\tau],\, \mathbb{E}_{x_0|x_\tau}[\mathbb{E}[x_0 \mid x_\tau] - x_0]\rangle\right] \\ &= \mathbb{E}_{x_\tau}\left[\langle g(x_\tau) - \mathbb{E}[x_0 \mid x_\tau],\, \mathbb{E}[x_0 \mid x_\tau] - \mathbb{E}[x_0 \mid x_\tau]\rangle\right] = 0. \end{aligned} \tag{272}$$

Since the Irreducible Error is independent of the model parameters g, minimizing the training objective L_{DM} is equivalent to minimizing the Estimation Error. Therefore, the optimal denoiser g^* is precisely the conditional expectation:

$$g^*(x_\tau) = \arg\min_g L_{\mathrm{DM}}(g) = \mathbb{E}[x_0 \mid x_\tau]. \tag{273}$$

Finally, by combining this result with Tweedie's formula (269), we establish a direct link between the denoiser and the score function:

$$s_\tau^*(x_\tau) = \frac{g^*(x_\tau) - x_\tau}{\tau^2}. \tag{274}$$

This identity confirms that training an MSE denoiser is mathematically equivalent to learning the score function. Consequently, we define the generalization error as the deviation from this optimal denoiser:

$$\mathcal{E}_{\mathrm{gen}}(g;\tau) = \|g(x_\tau) - g^*(x_\tau)\|_2, \tag{275}$$

which serves as a proxy for the error in the estimated score field required for sampling.

J.2. Score Function via Projection Decomposition

We consider the generative model

$$x_0 = \alpha_1 \mathbf{M}_1 z_1 + \alpha_2 \mathbf{M}_2 z_2, \tag{276}$$

where $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{d_1 \times d}$ have orthonormal columns and span two orthogonal feature subspaces. The noisy observation is

$$x_\tau = x_0 + \tau\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I_{d_1}). \tag{277}$$

Using the Tweedie identity, the score function is

$$s_\tau^*(x_\tau) = \frac{\mathbb{E}[x_0 \mid x_\tau] - x_\tau}{\tau^2}. \tag{278}$$

We first project the noisy sample onto the two feature subspaces. Define

$$y_1 = \mathbf{M}_1^\top x_\tau, \qquad y_2 = \mathbf{M}_2^\top x_\tau. \tag{279}$$

Using $x_\tau = \alpha_1 \mathbf{M}_1 z_1 + \alpha_2 \mathbf{M}_2 z_2 + \tau\epsilon$, we obtain

$$y_1 = \alpha_1 z_1 + \tau \mathbf{M}_1^\top \epsilon, \qquad y_2 = \alpha_2 z_2 + \tau \mathbf{M}_2^\top \epsilon. \tag{280}$$

Since $\mathbf{M}_1, \mathbf{M}_2$ have orthonormal columns and are orthogonal, the projected noise terms

$$\xi_1 = \mathbf{M}_1^\top \epsilon \sim \mathcal{N}(0, I_d), \qquad \xi_2 = \mathbf{M}_2^\top \epsilon \sim \mathcal{N}(0, I_d), \tag{281}$$

are i.i.d. standard Gaussian, and ξ_1, ξ_2 are independent. Hence we can write

$$y_1 = \alpha_1 z_1 + \tau\xi_1, \qquad y_2 = \alpha_2 z_2 + \tau\xi_2. \tag{282}$$

Because the prior and the Gaussian likelihood factorize across the two orthogonal subspaces, the posterior mean also decomposes as

$$\mathbb{E}[x_0 \mid x_\tau] = \alpha_1 \mathbf{M}_1 \hat{z}_1(y_1) + \alpha_2 \mathbf{M}_2 \hat{z}_2(y_2), \tag{283}$$

where

$$\hat{z}_k(y_k) = \mathbb{E}[z_k \mid y_k], \qquad k \in \{1, 2\}. \tag{284}$$

In the subsequent regime-wise analysis, it therefore suffices to control the behavior of the coordinate-wise denoisers $\hat{z}_k^j(y_k^j)$.

J.3. Posterior Mean under a Sparse Prior

Let $z = (z^1, \ldots, z^d) \in \{-1, 0, 1\}^d$ be a sparse latent vector whose coordinates are independent and satisfy

$$\Pr(z^j = 1) = \Theta(1/d), \qquad \Pr(z^j = -1) = \Theta(1/d), \qquad \Pr(z^j = 0) = 1 - \Theta(1/d). \tag{285}$$

Thus only $O(1)$ coordinates of z are nonzero in expectation.

The observation model is

$$y = \alpha z + \tau\xi, \qquad \xi \sim \mathcal{N}(0, I_d), \tag{286}$$

so that each coordinate satisfies

$$y^j = \alpha z^j + \tau\xi^j, \qquad \xi^j \sim \mathcal{N}(0, 1). \tag{287}$$

Since both the likelihood and the prior factorize across coordinates, the posterior also factorizes:

$$\mathbb{E}[z \mid y] = \big(\hat{z}^1(y^1), \ldots, \hat{z}^d(y^d)\big), \tag{288}$$

where each scalar posterior mean is

$$\hat{z}^j(y^j) = \mathbb{E}[z^j \mid y^j] = \Pr(z^j = 1 \mid y^j) - \Pr(z^j = -1 \mid y^j). \tag{289}$$

We now derive the scalar expression for a generic coordinate j. The likelihood is

$$p(y^j \mid z^j) \propto \exp\!\left(-\frac{(y^j - \alpha z^j)^2}{2\tau^2}\right), \tag{290}$$

and the prior has masses $\Theta(1/d)$ at $z^j = \pm 1$ and $1 - \Theta(1/d)$ at $z^j = 0$. Bayes' rule gives

$$\hat{z}^j(y^j) = \frac{p(y^j \mid 1)\,\Theta(1/d) - p(y^j \mid -1)\,\Theta(1/d)}{p(y^j \mid 0)\big(1 - \Theta(1/d)\big) + p(y^j \mid 1)\,\Theta(1/d) + p(y^j \mid -1)\,\Theta(1/d)}. \tag{291}$$

Introduce the likelihood ratios

$$L_1^j = \exp\!\left(\frac{2\alpha y^j - \alpha^2}{2\tau^2}\right), \qquad L_{-1}^j = \exp\!\left(\frac{-2\alpha y^j - \alpha^2}{2\tau^2}\right), \tag{292}$$

and define the sparsity bias

$$B = \ln \frac{1 - \Theta(1/d)}{\Theta(1/d)} = \Theta(\ln d). \tag{293}$$

Dividing numerator and denominator by $p(y^j \mid 0)(1 - \Theta(1/d))$ yields the closed-form scalar denoiser

$$\hat{z}^j(y^j) = \frac{e^{-B}\,(L_1^j - L_{-1}^j)}{1 + e^{-B}\,(L_1^j + L_{-1}^j)}. \tag{294}$$

Using the identities

$$L_1^j = \exp\left(-\frac{\alpha^2}{2\tau^2}\right)\exp\left(\frac{\alpha y^j}{\tau^2}\right), \qquad L_{-1}^j = \exp\left(-\frac{\alpha^2}{2\tau^2}\right)\exp\left(-\frac{\alpha y^j}{\tau^2}\right), \tag{295}$$

we obtain

$$L_1^j - L_{-1}^j = \exp\left(-\frac{\alpha^2}{2\tau^2}\right)\left(e^{\alpha y^j/\tau^2} - e^{-\alpha y^j/\tau^2}\right) = \exp\left(-\frac{\alpha^2}{2\tau^2}\right) 2\sinh\left(\frac{\alpha y^j}{\tau^2}\right), \tag{296}$$

and

$$L_1^j + L_{-1}^j = \exp\left(-\frac{\alpha^2}{2\tau^2}\right)\left(e^{\alpha y^j/\tau^2} + e^{-\alpha y^j/\tau^2}\right) = \exp\left(-\frac{\alpha^2}{2\tau^2}\right) 2\cosh\left(\frac{\alpha y^j}{\tau^2}\right). \tag{297}$$

Substituting these expressions gives the compact hyperbolic form

$$\hat{z}^j(y^j) = \frac{e^{-B}e^{-\frac{\alpha^2}{2\tau^2}} 2\sinh\left(\frac{\alpha y^j}{\tau^2}\right)}{1 + e^{-B}e^{-\frac{\alpha^2}{2\tau^2}} 2\cosh\left(\frac{\alpha y^j}{\tau^2}\right)}. \tag{298}$$

This expression admits a simple asymptotic interpretation. Define the logistic function

$$\sigma(u) = \frac{1}{1+e^{-u}}. \tag{299}$$

For large positive y^j, one has $L_{-1}^j \ll L_1^j$, so

$$\hat{z}^j(y^j) \approx \frac{e^{-B}L_1^j}{1 + e^{-B}L_1^j} = \sigma\Big(\log L_1^j - B\Big) = \sigma\left(\frac{2\alpha y^j - \alpha^2}{2\tau^2} - B\right), \tag{300}$$

which approaches 1 as y^j grows. For large negative y^j, one has $L_1^j \ll L_{-1}^j$, and

$$\hat{z}^j(y^j) \approx -\frac{e^{-B}L_{-1}^j}{1 + e^{-B}L_{-1}^j} = -\sigma\Big(\log L_{-1}^j - B\Big) = -\sigma\left(\frac{-2\alpha y^j - \alpha^2}{2\tau^2} - B\right), \tag{301}$$

which approaches -1 as y^j decreases. When y^j is close to zero, the contributions of L_1^j and L_{-1}^j nearly cancel in the numerator while the denominator is dominated by the constant term, so $\hat{z}^j(y^j)$ stays close to zero. Thus the ternary posterior mean interpolates between -1, 0, and 1 via a symmetric soft-thresholding nonlinearity.

Having established the behavior of the scalar posterior mean, we now substitute this estimator into the structured score decomposition and work out its regime-wise behavior in detail.

High-noise regime. Assume $\tau \geq (1+\delta)\frac{\alpha_1}{\sqrt{2\ln d}}$ for some fixed $\delta > 0$. For any coordinate j and subspace $k \in \{1,2\}$, the observation satisfies

$$y_k^j = \alpha_k z_k^j + \tau\xi_k^j, \qquad \xi_k^j \sim \mathcal{N}(0,1). \tag{302}$$

In this regime, the signal-to-noise ratio is insufficient to overcome the sparsity bias $B = \Theta(\ln d)$, and the scalar posterior mean remains strongly suppressed.

We now bound the noise-averaged denoiser. Fix $k \in \{1,2\}$ and $j \in [d]$. Using the positive-tail approximation (which applies whenever the likelihood ratio is dominated by L_1^j),

$$\hat{z}^j(y_k^j) \approx \sigma\left(\frac{2\alpha_k y_k^j - \alpha_k^2}{2\tau^2} - B\right) = \sigma\Big(\mu_k^j + \frac{\alpha_k}{\tau}\,\xi_k^j\Big), \tag{303}$$

where

$$\mu_k^j = \frac{\alpha_k^2(2z_k^j - 1)}{2\tau^2} - B. \tag{304}$$

Similarly, when the negative tail dominates,

$$-\hat{z}^j(y_k^j) \approx \sigma\Big(\tilde{\mu}_k^j + \frac{\alpha_k}{\tau}\,\xi_k^j\Big), \qquad \tilde{\mu}_k^j = \frac{\alpha_k^2(-2z_k^j - 1)}{2\tau^2} - B. \tag{305}$$

In both cases, the argument of the sigmoid is an affine Gaussian function of ξ_k^j.

Applying Lemma 5 with $X = \mu_k^j + (\alpha_k/\tau)\xi_k^j \sim \mathcal{N}\Big(\mu_k^j, \frac{\alpha_k^2}{\tau^2}\Big)$, and using the elementary inequality $\sigma(x) \le e^x$ for all $x \in \mathbb{R}$, together with the Gaussian moment generating function, we obtain

$$\mathbb{E}_{\xi_k^j}\big[\sigma(X)\big] \le \exp\!\left(\mu_k^j + \frac{\alpha_k^2}{2\tau^2}\right), \qquad \mathbb{E}_{\xi_k^j}\big[\sigma(\tilde{X})\big] \le \exp\!\left(\tilde{\mu}_k^j + \frac{\alpha_k^2}{2\tau^2}\right), \tag{306}$$

where $\tilde{X} = \tilde{\mu}_k^j + (\alpha_k/\tau)\xi_k^j$.

Under the assumption $\tau \ge (1+\delta)\alpha_1/\sqrt{2\ln d}$, we have

$$\frac{\alpha_k^2}{2\tau^2} \le \frac{1}{(1+\delta)^2}\,\ln d \quad \text{and} \quad B = c_0 \ln d$$

for some absolute constant $c_0 > 1$. Consequently, there exists $c(\delta) > 0$ such that

$$\mu_k^j + \frac{\alpha_k^2}{2\tau^2} \le -c(\delta)\,\ln d \quad \text{and} \quad \tilde{\mu}_k^j + \frac{\alpha_k^2}{2\tau^2} \le -c(\delta)\,\ln d$$

uniformly over all $j \in [d]$ and $k \in \{1,2\}$. Therefore,

$$\Big|\mathbb{E}_{\xi_k^j}\Big[\hat{z}^j(y_k^j) \mid z_k^j\Big]\Big| \le C\,d^{-c(\delta)} \tag{307}$$

for some constant $C > 0$.

Hence, in the high-noise regime, the conditional expectation of each coordinate-wise posterior mean is uniformly negligible across both feature subspaces. This noise-averaged suppression is the only property required for the subsequent comparison between $g(x_\tau)$ and $\mathbb{E}[x_0 \mid x_\tau]$.

Intermediate-noise regime. Assume $(1+\delta)\alpha_2/\sqrt{2\ln d} \ll \tau \ll (1-\delta)\alpha_1/\sqrt{2\ln d}$. We analyze the noise-averaged behavior of the coordinate-wise posterior mean $\hat{z}^j(y_k^j)$ for each subspace $k \in \{1,2\}$.

Strong component ($k = 1$). For an active coordinate with $z_1^j = 1$, the observation satisfies

$$y_1^j = \alpha_1 + \tau\xi_1^j, \qquad \xi_1^j \sim \mathcal{N}(0,1). \tag{308}$$

Using the large-positive approximation of the scalar posterior mean,

$$\hat{z}^j(y_1^j) \approx \sigma\left(\frac{2\alpha_1 y_1^j - \alpha_1^2}{2\tau^2} - B\right) = \sigma\Big(\mu_1^+ + \frac{\alpha_1}{\tau}\xi_1^j\Big), \tag{309}$$

where

$$\mu_1^+ = \frac{\alpha_1^2}{2\tau^2} - B. \tag{310}$$

Under the assumption $\tau \le (1-\delta)\alpha_1/\sqrt{2\ln d}$, we have

$$\frac{\alpha_1^2}{2\tau^2} \ge \frac{1}{(1-\delta)^2}\ln d, \qquad B = c_0 \ln d$$

for some absolute constant $c_0 > 1$. Hence there exists $c = c(\delta) > 0$ such that $\mu_1^+ \ge c\ln d$.

Applying Lemma 5 with $X = \mu_1^+ + (\alpha_1/\tau)\xi_1^j$, and using the inequality $1 - \sigma(x) \le e^{-x}$, together with the Gaussian moment generating function, we obtain

$$\mathbb{E}_{\xi_1^j}\Big[1 - \hat{z}^j(y_1^j) \mid z_1^j = 1\Big] \le \exp\!\left(-\mu_1^+ + \frac{\alpha_1^2}{2\tau^2}\right) \le C\, d^{-c}, \tag{311}$$

which implies

$$\mathbb{E}_{\xi_1^j}\Big[\hat{z}^j(y_1^j) \mid z_1^j = 1\Big] = 1 - O(d^{-c}). \tag{312}$$

For inactive coordinates with $z_1^j = 0$, we have $y_1^j = \tau\xi_1^j$ and

$$\hat{z}^j(y_1^j) \approx \sigma\Big(\mu_1^0 + \frac{\alpha_1}{\tau}\xi_1^j\Big), \qquad \mu_1^0 = -\frac{\alpha_1^2}{2\tau^2} - B. \tag{313}$$

Since $\mu_1^0 \le -c\ln d$, the inequality $\sigma(x) \le e^x$ yields

$$\Big|\mathbb{E}_{\xi_1^j}\Big[\hat{z}^j(y_1^j) \mid z_1^j = 0\Big]\Big| \le C\, d^{-c}. \tag{314}$$

Thus, after averaging over the Gaussian noise, the strong component is recovered coordinate-wise up to vanishing error.

Weak component ($k = 2$). For any coordinate j, the observation satisfies

$$y_2^j = \alpha_2 z_2^j + \tau\xi_2^j. \tag{315}$$

Using the same approximation,

$$\hat{z}^j(y_2^j) \approx \sigma\Big(\mu_2^j + \frac{\alpha_2}{\tau}\xi_2^j\Big), \qquad \mu_2^j = \frac{\alpha_2^2(2z_2^j - 1)}{2\tau^2} - B. \tag{316}$$

The assumption $\tau \ge (1+\delta)\alpha_2/\sqrt{2\ln d}$ implies

$$\frac{\alpha_2^2}{2\tau^2} \le \frac{1}{(1+\delta)^2}\ln d, \qquad B = c_0 \ln d,$$

so there exists $c' > 0$ such that $\mu_2^j \le -c'\ln d$ uniformly over all j. Applying Lemma 5 and $\sigma(x) \le e^x$, we obtain

$$\Big|\mathbb{E}_{\xi_2^j}\Big[\hat{z}^j(y_2^j) \mid z_2^j\Big]\Big| \le C\, d^{-c'} \qquad \forall j. \tag{317}$$

Conclusion. In the intermediate-noise regime, the noise-averaged posterior mean satisfies

$$\mathbb{E}_\xi\Big[\hat{z}_1^j(y_1^j) \mid z_1^j\Big] = z_1^j + O(d^{-c}), \qquad \mathbb{E}_\xi\Big[\hat{z}_2^j(y_2^j) \mid z_2^j\Big] = O(d^{-c'}), \tag{318}$$

uniformly over coordinates. Thus, after averaging over the Gaussian noise, the strong component is reliably recovered while the weak component remains suppressed.

Low-noise regime. Assume $\tau \le (1-\delta)\frac{\alpha_2}{\sqrt{2\ln d}}$ for some $\delta > 0$. We analyze the noise-averaged behavior of the coordinate-wise posterior mean $\hat{z}^j(y_k^j)$ for both subspaces $k \in \{1, 2\}$.

Weak component ($k = 2$). For an active coordinate with $z_2^j = 1$, the observation is

$$y_2^j = \alpha_2 + \tau\xi_2^j, \qquad \xi_2^j \sim \mathcal{N}(0,1). \tag{319}$$

Using the large-positive approximation of the scalar posterior mean,

$$\hat{z}^j(y_2^j) \approx \sigma\left(\frac{2\alpha_2 y_2^j - \alpha_2^2}{2\tau^2} - B\right) = \sigma\Big(\mu_2^+ + \frac{\alpha_2}{\tau}\xi_2^j\Big), \tag{320}$$

where

$$\mu_2^+ = \frac{\alpha_2^2}{2\tau^2} - B. \tag{321}$$

Under the assumption $\tau \le (1-\delta)\alpha_2/\sqrt{2\ln d}$, we have

$$\frac{\alpha_2^2}{2\tau^2} \ge \frac{1}{(1-\delta)^2}\ln d, \qquad B = c_0 \ln d$$

for some absolute constant $c_0 > 1$. Hence there exists $c = c(\delta) > 0$ such that $\mu_2^+ \ge c \ln d$.

Applying Lemma 5 with $X = \mu_2^+ + (\alpha_2/\tau)\xi_2^j$, and using the inequality $1 - \sigma(x) \le e^{-x}$, together with the Gaussian moment generating function, we obtain

$$\mathbb{E}_{\xi_2^j}\Big[1 - \hat{z}^j(y_2^j) \mid z_2^j = 1\Big] \le \exp\!\left(-\mu_2^+ + \frac{\alpha_2^2}{2\tau^2}\right) \le C\, d^{-c}, \tag{322}$$

which implies

$$\mathbb{E}_{\xi_2^j}\Big[\hat{z}^j(y_2^j) \mid z_2^j = 1\Big] = 1 - O(d^{-c}). \tag{323}$$

For inactive coordinates with $z_2^j = 0$, the observation reduces to $y_2^j = \tau\xi_2^j$, and the logistic argument has mean

$$\mu_2^0 = -\frac{\alpha_2^2}{2\tau^2} - B \le -c\ln d. \tag{324}$$

Using $\sigma(x) \le e^x$ and Lemma 5, we obtain

$$\left|\mathbb{E}_{\xi_2^j}\Big[\hat{z}^j(y_2^j) \mid z_2^j = 0\Big]\right| \le C\, d^{-c}. \tag{325}$$

Strong component ($k = 1$). Since $\alpha_1 > \alpha_2$, the condition $\tau \le (1-\delta)\alpha_2/\sqrt{2\ln d}$ implies $\tau \ll \alpha_1/\sqrt{2\ln d}$. Repeating the same argument as in the intermediate-noise regime yields

$$\mathbb{E}_{\xi_1^j}\Big[\hat{z}^j(y_1^j) \mid z_1^j\Big] = z_1^j + O(d^{-c}), \tag{326}$$

uniformly over all coordinates.

Conclusion. In the low-noise regime, the noise-averaged posterior mean satisfies

$$\mathbb{E}_\xi\Big[\hat{z}_1^j(y_1^j) \mid z_1^j\Big] = z_1^j + O(d^{-c}), \qquad \mathbb{E}_\xi\Big[\hat{z}_2^j(y_2^j) \mid z_2^j\Big] = z_2^j + O(d^{-c}), \tag{327}$$

uniformly over coordinates. Thus, after averaging over the Gaussian noise, both the strong and weak components are recovered with vanishing error, which completes the three-regime characterization of the posterior mean dynamics.

Summary. The above analysis establishes a sharp three-regime characterization of the noise-averaged posterior mean under a ternary sparse prior. In the high-noise regime, the sparsity bias dominates the likelihood uniformly across subspaces, and the conditional expectation $\mathbb{E}_\xi[\hat{z}^j(y_k^j) \mid z_k^j]$ is exponentially suppressed for all coordinates, yielding a purely shrinkage-dominated behavior. In the intermediate-noise regime, the posterior mean exhibits a separation of scales: coordinates associated with the strong component are recovered with vanishing error after noise averaging, while those associated with the weak component remain suppressed. In the low-noise regime, the signal-to-noise ratio exceeds the sparsity threshold for both components, and the noise-averaged posterior mean converges coordinate-wise to the ground-truth latent vector. Together, these results show that the ternary posterior mean undergoes a well-defined phase transition structure governed by the relative magnitude of τ and $\alpha_k/\sqrt{2\ln d}$, providing a rigorous foundation for the subsequent comparison between learned score functions and the exact conditional expectation $\mathbb{E}[x_0 \mid x_\tau]$.

J.4. Directional Generalization Error

We define the subspace-wise directional generalization error to quantify the mismatch between the learned denoiser g and the oracle denoiser g^* in each feature direction:

$$\mathcal{E}_k(g;\tau) = \mathbb{E}\Big[\|\mathbf{M}_k^\top g(x_\tau) - \mathbf{M}_k^\top g^*(x_\tau)\|_2^2\Big], \qquad k \in \{1,2\}. \tag{328}$$

The oracle denoiser satisfies

$$g^*(x_\tau) = \mathbb{E}[x_0 \mid x_\tau] = \alpha_1 \mathbf{M}_1 \hat{z}_1(y_1) + \alpha_2 \mathbf{M}_2 \hat{z}_2(y_2), \tag{329}$$

where $\hat{z}_k(y_k)$ is the coordinate-wise posterior mean. The total generalization error decomposes as:

$$\mathcal{E}_{\text{gen}}^2(g;\tau) = \|g(x_\tau) - g^*(x_\tau)\|_2^2 = \mathcal{E}_1(g;\tau) + \mathcal{E}_2(g;\tau) + \mathcal{E}_\perp(g;\tau), \tag{330}$$

where $\mathcal{E}_\perp$ captures error orthogonal to both subspaces.

J.5. High-Noise Regime Analysis

We now analyze the high-noise regime where $\tau \in [\tau_1, \tau_{\max}]$ with $\tau_1 = \Theta(\alpha_2/\sqrt{\log d})$ and $\tau_{\max} = \Theta(\alpha_1/\sqrt{\log d})$. In this regime, we show that both g_{joint} and g_{std} match the Bayes-optimal denoiser g^*, but the reconstruction error remains $\Theta(1/d^{2c_0})$ due to the irreducible error from unrecovered $\mathbf{M}_2$ features.

We begin by characterizing the Bayes-optimal behavior. In the high-noise regime, the noise level satisfies $(1+\delta)\alpha_2/\sqrt{2\ln d} \ll \tau \ll (1-\delta)\alpha_1/\sqrt{2\ln d}$ for some $\delta > 0$. By the posterior mean analysis (Section D), the noise-averaged posterior means satisfy:

$$\mathbb{E}_\xi\Big[\hat{z}_1^j(y_1^j) \mid z_1^j\Big] = z_1^j + O(d^{-c}), \qquad \mathbb{E}_\xi\Big[\hat{z}_2^j(y_2^j) \mid z_2^j\Big] = O(d^{-c'}), \tag{331}$$

uniformly over coordinates. This means the $\mathbf{M}_1$ component is recoverable while the $\mathbf{M}_2$ component is suppressed by noise.

Analysis of g_{joint}. Under joint denoising-sparsity scheduling, neurons in $S_{1j,\text{sure}}$ achieve pure $\mathbf{M}_{1j}$ alignment by Stage I (Theorem 1):

$$|\langle w_i^{(T_1)}, \mathbf{M}_{1j}\rangle|^2 \geq (1-o(1))\|w_i^{(T_1)}\|_2^2, \qquad |\langle w_i^{(T_1)}, \mathbf{M}_{2k}\rangle|^2 = o(1)\|w_i^{(T_1)}\|_2^2 \ \ \forall k. \tag{332}$$

For the gating decision, since $\tau \ll \alpha_1/\sqrt{2\ln d}$, Lemma 4 yields:

$$\Pr(w_i^\top x_\tau \geq b_i \mid z_1^j = 1) = 1 - d^{-\Omega(1)}, \qquad \Pr(w_i^\top x_\tau \geq b_i \mid z_1^j = 0) = d^{-\Omega(1)}. \tag{333}$$

The model-induced latent estimate $\tilde{z}_1^j$ thus satisfies:

$$\mathbb{E}_\xi\Big[\tilde{z}_1^j(y_1^j) \mid z_1^j\Big] = z_1^j + O(d^{-c}), \tag{334}$$

matching the Bayes posterior behavior. For $\mathbf{M}_2$ directions, the noise level exceeds α_2, so gating probabilities satisfy $\Pr(w_i^\top x_\tau \geq b_i \mid z_2^j) = \frac{1}{2} \pm d^{-\Omega(1)}$, yielding:

$$\mathbb{E}_\xi\Big[\tilde{z}_2^j(y_2^j) \mid z_2^j\Big] = O(d^{-c'}). \tag{335}$$

Combining these, we obtain:

$$\mathcal{E}_1(g_{\text{joint}};\tau) = o(1), \qquad \mathcal{E}_2(g_{\text{joint}};\tau) = o(1). \tag{336}$$

Analysis of g_{std}. Under standard training, all neurons participate from the beginning. By Theorem 1, neurons in $S_{1j,\text{sure}}$ still achieve $\mathbf{M}_{1j}$ alignment (though potentially with some $\mathbf{M}_1$ cross-contamination). The key observation is that in the high-noise regime, the $\mathbf{M}_2$ signal is overwhelmed by noise for both training protocols. Specifically, the signal-to-noise ratio for $\mathbf{M}_2$ features satisfies:

$$\frac{\alpha_2|\langle w_i, \mathbf{M}_{2j}\rangle|}{\tau\|w_i\|_2/\sqrt{d_1}} = O\left(\frac{\alpha_2\sqrt{d_1}}{\tau_{\max}}\right) = O\left(\frac{\alpha_2\sqrt{d_1\log d}}{\alpha_1}\right) = o(1), \tag{337}$$

since $\alpha_2 = \Theta(1/d^{c_0})$ and $\alpha_1 = \Theta(1)$.

Since both the Bayes-optimal denoiser and the learned denoiser g_{std} produce negligible output along $\mathbf{M}_2$ directions (due to noise suppression), we have:

$$\mathcal{E}_1(g_{\text{std}}; \tau) = o(1), \qquad \mathcal{E}_2(g_{\text{std}}; \tau) = o(1). \tag{338}$$

Irreducible error from unrecovered $\mathbf{M}_2$. While both learned denoisers match the Bayes-optimal in terms of the directional error $\mathcal{E}_k$, the oracle itself cannot recover the $\mathbf{M}_2$ component in the high-noise regime. By the bias-variance decomposition:

$$\mathbb{E}[\|g(x_\tau) - x_0\|_2^2] = \underbrace{\mathbb{E}[\|g - g^*\|_2^2]}_{\text{estimation error}} + \underbrace{\mathbb{E}[\|g^* - x_0\|_2^2]}_{\text{irreducible error}}. \tag{339}$$

In the high-noise regime, the posterior mean for $\mathbf{M}_2$ coordinates satisfies $\mathbb{E}_\xi[\hat{z}_2^j \mid z_2^j] = O(d^{-c})$ (Eq. (331)), so the oracle output along $\mathbf{M}_2$ is negligible. The irreducible error is therefore:

$$\mathbb{E}[\|g^*(x_\tau) - x_0\|_2^2] \geq \mathbb{E}[\|\mathbf{M}_2^\top (g^* - x_0)\|_2^2] = \mathbb{E}[\|\alpha_2 z_2 - O(d^{-c})\|_2^2] = \Theta(\alpha_2^2) = \Theta(1/d^{2c_0}). \tag{340}$$

Conclusion for high-noise regime. Both training protocols match the Bayes-optimal denoiser in terms of directional error:

$$\mathcal{E}_1(g; \tau) + \mathcal{E}_2(g; \tau) = \|g - g^*\|_2^2 = o(1). \tag{341}$$

However, the total reconstruction error is dominated by the irreducible component from unrecovered $\mathbf{M}_2$ features:

$$\mathbb{E}[\|g_{\text{joint}}(x_\tau) - x_0\|_2^2] = \Theta(1/d^{2c_0}), \quad \mathbb{E}[\|g_{\text{std}}(x_\tau) - x_0\|_2^2] = \Theta(1/d^{2c_0}), \quad \tau \in [\tau_1, \tau_{\max}]. \tag{342}$$

This matches Theorem 5(a): in the high-noise regime, both methods achieve the same generalization accuracy $\Theta(1/d^{2c_0})$, with the error arising entirely from the information-theoretically unrecoverable $\mathbf{M}_2$ component.

J.6. Low-Noise Regime Analysis

We now analyze the low-noise regime where $\tau \in [\tau_{\min}, \tau_1]$ with $\tau_{\min} = \Theta(1/d_1)$. In this regime, the two training protocols exhibit fundamentally different generalization behavior.

Bayes-optimal behavior. In the low-noise regime, the noise level satisfies $\tau \leq (1-\delta)\alpha_2/\sqrt{2\ln d}$. By the posterior mean analysis, the Bayes-optimal denoiser achieves full recovery of both components:

$$\mathbb{E}_\xi\left[\hat{z}_1^j(y_1^j) \mid z_1^j\right] = z_1^j + O(d^{-c}), \qquad \mathbb{E}_\xi\left[\hat{z}_2^j(y_2^j) \mid z_2^j\right] = z_2^j + O(d^{-c}). \tag{343}$$

Analysis of g_{joint} (Pure $\mathbf{M}_2$ Alignment). Under joint denoising-sparsity scheduling, neurons in $S_{2j,\text{sure}}$ are frozen during Stage I and only trained during Stage II with low noise. By Theorem 2, these neurons achieve **pure** $\mathbf{M}_{2j}$ alignment:

$$|\langle \hat{w}_i^{(T_1+T_2)}, \mathbf{M}_{2j} \rangle|^2 = (1 - o(1))\, \|\hat{w}_i^{(T_1+T_2)}\|_2^2, \tag{344}$$

with negligible $\mathbf{M}_1$ contamination:

$$|\langle \hat{w}_i^{(T_1+T_2)}, \mathbf{M}_{1j'} \rangle| = O\left(\frac{\sigma_0}{\sqrt{d}}\right), \qquad \forall j' \in [d]. \tag{345}$$

For the gating decision in the low-noise regime, since the neurons have pure $\mathbf{M}_{2j}$ alignment and $\tau < \alpha_2/\sqrt{2\ln d}$, Lemma 4 yields:

$$\Pr(\hat{w}_i^\top x_\tau \geq \hat{b}_i \mid z_2^j = 1) = 1 - d^{-\Omega(1)}, \qquad \Pr(\hat{w}_i^\top x_\tau \geq \hat{b}_i \mid z_2^j = 0) = d^{-\Omega(1)}. \tag{346}$$

The model-induced latent estimate thus satisfies:

$$\mathbb{E}_\xi\left[\hat{z}_2^j(y_2^j) \mid z_2^j\right] = z_2^j + O(d^{-c}), \tag{347}$$

matching the Bayes posterior behavior (343). Combined with the $\mathbf{M}_1$ recovery from Stage I, we obtain:

$$\mathcal{E}_1(g_{\text{joint}};\tau) = o(1), \qquad \mathcal{E}_2(g_{\text{joint}};\tau) = o(1). \tag{348}$$

Analysis of g_{std} (Entangled $\mathbf{M}_2$ Alignment). Under standard training, neurons are exposed to the full noise range $[\tau_{\min}, \tau_{\max}]$ throughout training, including high-noise samples where $\tau > \alpha_2/\sqrt{\log d}$. By Theorem 4, neurons that learn $\mathbf{M}_2$ features exhibit **entangled** representations:

$$\bar{w}_i^{(t)} = \sum_{j\in\mathcal{N}_i} \bar{\beta}_{i,j}\mathbf{M}_{2j} + \bar{\mathbf{r}}_i, \tag{349}$$

where $|\mathcal{N}_i| = \Theta(\sqrt{d})$ directions are learned simultaneously, and the residual satisfies $\|\bar{\mathbf{r}}_i\|_2^2 = \Theta(\|\bar{w}_i^{(t)}\|_2^2)$.

This entanglement arises because high-noise samples trigger the Mixed Regime of Lemma 4, destroying gating separation. When the noise dominates the $\mathbf{M}_2$ signal, the gate activates with probability $\approx 1/2$ regardless of which $\mathbf{M}_{2j}$ feature is present, causing multiple directions to receive concurrent gradient updates.

Consequently, neuron i fires whenever **any** feature in $\mathcal{N}_i$ is present. When $z_2^k = 0$ but $z_2^j = 1$ for some $j \in \mathcal{N}_i$ with $j \neq k$, the entangled alignment produces erroneous activation:

$$\mathbb{E}_\xi\left[\tilde{z}_2^k(y_2^k) \mid z_2^k = 0, \exists j \in \mathcal{N}_i : z_2^j = 1\right] = \Theta(1). \tag{350}$$

This demonstrates that the model leaks features across $\mathbf{M}_2$ directions during denoising.

Comparing with the Bayes-optimal behavior (343), we obtain a non-vanishing gap:

$$\mathcal{E}_2(g_{\text{std}};\tau) = \|\tilde{z}_2(y_2) - \hat{z}_2(y_2)\|_2 = \Omega(1). \tag{351}$$

In contrast, the $\mathbf{M}_1$ component remains correctly captured since neurons in $S_{1j,\text{sure}}$ achieve pure $\mathbf{M}_{1j}$ alignment:

$$\mathcal{E}_1(g_{\text{std}};\tau) = o(1). \tag{352}$$

Conclusion for low-noise regime. The two training protocols exhibit a separation:

$$\mathcal{E}_{\text{gen}}(g_{\text{joint}};\tau) = o(1), \qquad \mathcal{E}_{\text{gen}}(g_{\text{std}};\tau) = \Theta(1), \qquad \tau \in [\tau_{\min}, \tau_1]. \tag{353}$$

J.7. Generalization Summary

We summarize the generalization behavior across noise regimes:

	High-noise $[\tau_1, \tau_{\max}]$	Low-noise $[\tau_{\min}, \tau_1]$
g_{joint}	$\mathcal{E}_{\text{gen}} = o(1)$	$\mathcal{E}_{\text{gen}} = o(1)$
g_{std}	$\mathcal{E}_{\text{gen}} = o(1)$	$\mathcal{E}_{\text{gen}} = \Theta(1)$
$\mathbf{M}_1$ recovery	Both succeed	Both succeed
$\mathbf{M}_2$ recovery	Both suppressed (noise)	Joint: pure; Std: entangled

The key insight is that the generalization gap between the two protocols manifests **exclusively** in the low-noise regime. In the high-noise regime, the $\mathbf{M}_2$ component is suppressed by noise for both the Bayes-optimal denoiser and the learned denoisers, so both training protocols achieve near-optimal performance.

The failure of standard training in the low-noise regime arises from the entangled $\mathbf{M}_2$ alignment established in Theorem 4: neurons that encode $\mathbf{M}_2$ features are not pure single-direction detectors but dense mixtures over multiple $\mathbf{M}_{2j}$ atoms. This entanglement causes systematic errors when the network attempts to recover $\mathbf{M}_2$ components from low-noise observations.

In contrast, joint denoising-sparsity scheduling (Theorem 2) ensures that neurons achieve pure $\mathbf{M}_{2j}$-alignment for each $j \in [d]$, enabling accurate recovery of both $\mathbf{M}_1$ and $\mathbf{M}_2$ components across all noise levels.

This completes the proof of Theorem 5.