WHEN FREQUENCIES INTERACT: A GRADIENT VIEW OF DIFFUSION TRAINING

Anonymous authors
Paper under double-blind review

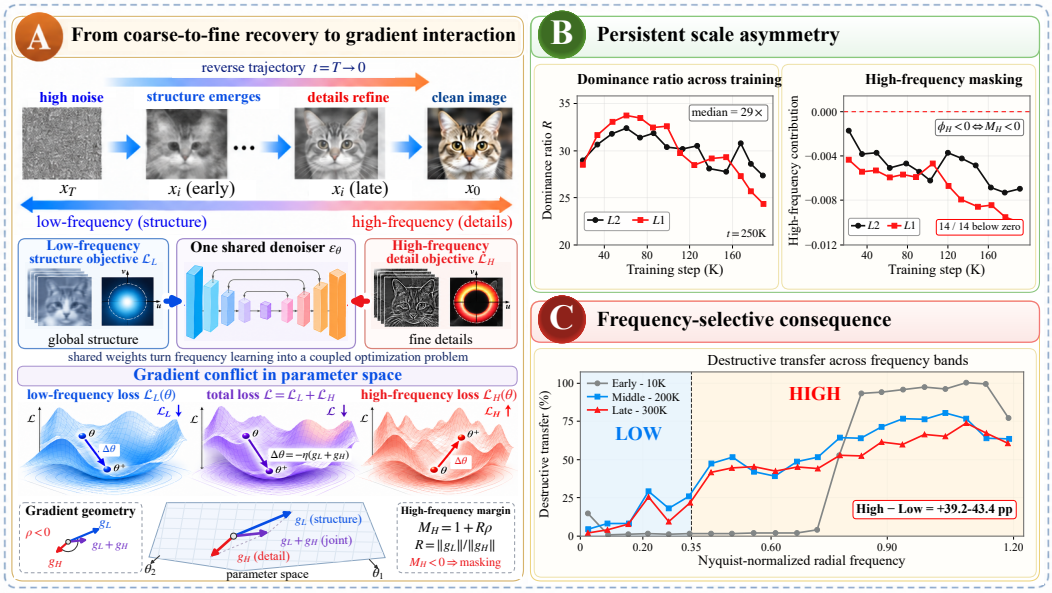

Figure 1: Frequency interaction in a shared denoiser. (A) Band objectives share parameters; $M_H = 1 + R\rho$ tests whether a small joint update increases high-frequency error. (B) At diffusion timestep $t = 250$, the input block has median gradient norm ratio R near 29 and negative high-frequency contribution ϕ_H at all 14 post-warmup checkpoints under L2 and L1. (C) Low-frequency test updates increase error more often above $c = 0.35$ across training stages.

ABSTRACT

Diffusion models typically learn low-frequency structure before high-frequency detail. Existing explanations focus on data variance and differences in learning speed across frequencies. We study frequency learning from a gradient perspective and find that different bands can influence one another through shared denoiser parameters. Our analysis shows that orthogonal frequency residuals can produce opposing gradients, with their magnitudes and directions determining how each band is affected. Low-frequency gradient dominance can amplify even mild opposition, allowing a sufficiently small joint update to reduce total error while increasing high-frequency error. Measurements and test updates on a standard denoiser reveal this interaction and its dependence on noise level. Motivated by these findings, we propose SpecDiffusion, whose Frequency-Aware Loss adjusts the relative band weights across diffusion timesteps, optionally complemented by a lightweight frequency-view adapter. Across pixel- and latent-space diffusion models with convolutional and Transformer denoisers, SpecDiffusion consistently improves matched baselines. These results offer a gradient-based view of frequency learning and connect it to a simple training objective.

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

1 INTRODUCTION

Denoising diffusion models generate images by recovering global structure and local detail from noisy inputs (Ho et al., 2020; Rombach et al., 2022; Peebles & Xie, 2023; Karras et al., 2022). Global structure is mainly associated with low frequencies, while edges and textures involve high-frequency components (Torralba & Oliva, 2003). Existing methods use frequency-domain supervision, wavelet decomposition, and band-wise modeling to treat these components separately (Jiang et al., 2021; Fuoli et al., 2021; Yang et al., 2025; Ma et al., 2025b). In a shared denoiser, the bands are learned through the same parameters, so an update for one band can affect the other.

Existing theory explains differences in learning speed through data statistics and network structure (Rahaman et al., 2019; Xu et al., 2020; Luo et al., 2021; Wang & Pehlevan, 2025). We ask how the same parameter update affects different bands. Shared parameters allow an update that reduces low-frequency error to change the high-frequency prediction. This raises a concrete question: *when a shared denoiser reduces its total error, must its high-frequency error also decrease?*

To analyze how one parameter update affects both bands, we examine the denoising error of a batch of noisy training samples at the same diffusion timestep. We split the squared error into low- and high-frequency terms. Their residuals are orthogonal in output space, but their gradients through shared parameters can oppose one another. When low-frequency interference outweighs the high-frequency gradient's own descent, a sufficiently small joint step reduces total error while increasing high-frequency error. We combine the gradient magnitudes and alignment to derive the condition for this first-order increase.

We compare a standard denoiser's input-layer gradients for both bands across training stages and noise levels. The joint update's effect on high-frequency error varies with noise level, with a first-order increase at some evaluated levels. Small test steps along the low-frequency descent direction also increase error more often in the high-frequency region, making the effect of low-frequency descent visible in the output error.

Loss weights change the relative gradient contributions. We derive ratios that allow both band errors to decrease locally, providing a basis for SpecDiffusion. Its core component, Frequency-Aware Loss (FAL), uses the noise level to set band priorities: low frequencies associated with global structure at high noise, and high frequencies associated with detail at low noise. Existing timestep-weighting methods adjust the relative importance of different noise levels during training (Choi et al., 2022; Hang et al., 2023; Salimans & Ho, 2022; Kingma et al., 2021); FAL additionally adjusts the relative band weights within each noise level. A lightweight input adapter supplies local low- and high-frequency views without changing the prediction head or sampler.

Our main contributions are: **(i)** We explain how frequency bands interact through shared parameters and characterize when this local interaction vanishes; **(ii)** we derive when total-error descent increases high-frequency error and examine this effect through gradient measurements and low-frequency test updates; **(iii)** we propose FAL based on the weighting analysis. Component experiments show that FAL alone reduces FID, while matched comparisons across pixel- and latent-space denoisers further support the effectiveness of SpecDiffusion.

2 RELATED WORK

Frequency-aware generative learning. Frequency-aware generative methods address spectral artifacts and differences in how networks learn frequency components (Schwarz et al., 2021). One approach changes the supervision: Focal Frequency Loss (Jiang et al., 2021), Fourier-space supervision (Fuoli et al., 2021), and complex-domain regularization (Gao & Cao, 2021) incorporate frequency information into the training objective. Another approach changes the representation or denoising process, through DCT generation (Ning et al., 2025), wavelet guidance (Yang et al., 2025), and frequency-decoupled pixel modeling (Ma et al., 2025b). Related developments include direct pixel diffusion (Chen et al., 2025) and frequency-aligned distillation (Choi et al., 2025). Blue-noise diffusion (Huang et al., 2024) and phase-preserving corruption (Zeng et al., 2025) modify the corruption process, while inference-time methods exploit coarse-to-fine denoising for conditional synthesis (Choi et al., 2021; Meng et al., 2022). These approaches use frequency structure to guide generation. We study how low- and high-frequency losses interact during joint training.

Spectral learning theory explains differences in learning rates through data statistics and network structure (Rahaman et al., 2019; Xu et al., 2020; Luo et al., 2021; Wang & Pehlevan, 2025). Our analysis examines how a joint update driven by the low- and high-frequency losses affects each component. The current denoiser Jacobian describes their coupling; gradient scale and alignment determine whether total-error descent increases high-frequency error to first order.

Gradient dynamics and diffusion weighting. Multi-task learning distinguishes scale imbalance from directional conflict (Sener & Koltun, 2018; Yu et al., 2020; Liu et al., 2021). Diffusion training can itself be treated as a family of denoising tasks and exhibits negative transfer between noise levels (Go et al., 2023; Ma et al., 2025a). These studies motivate examining both the scale and direction of gradients associated with different objectives. Here the objectives are frequency bands at the same diffusion timestep, and the question is whether their joint direction decreases each band's error.

Existing diffusion objectives address timestep imbalance with scalar weights (Choi et al., 2022; Hang et al., 2023; Kingma et al., 2021). At a fixed parameter state and timestep, multiplying both band losses by a common positive scalar leaves their gradient norm ratio, cosine similarity, and contribution signs unchanged. FAL changes the relative band weights within each timestep and can be combined with scalar timestep weighting.

3 Theory: Frequency Interaction under Shared Parameters

3.1 Band errors and parameter updates

Fix a diffusion timestep t and a minibatch of m samples. For a denoiser f_θ with parameters θ, let x_t be its noisy input and y_t a target such as noise or velocity. Stack the residuals $r_\theta = f_\theta(x_t, t) - y_t$ over samples and multiply by $1/\sqrt{m}$ to obtain $\mathbf{r}$, so its squared norm gives the minibatch-average error.

Let P_L, P_H be complementary orthogonal low- and high-frequency projections, and let $\Pi_B = \mathrm{diag}(P_B, \ldots, P_B)$ apply P_B to each sample. For a selected parameter block $\vartheta \subseteq \theta$, define

$$\mathbf{r}_B = \Pi_B \mathbf{r}, \qquad \mathcal{E}_B = \frac{1}{2}\|\mathbf{r}_B\|_2^2, \qquad g_B = \nabla_\vartheta \mathcal{E}_B, \qquad B \in \{L, H\}. \tag{1}$$

Orthogonality gives $\mathcal{E} = \mathcal{E}_L + \mathcal{E}_H$ and $\nabla_\vartheta \mathcal{E} = g_L + g_H$, but does not ensure that both band errors improve under a joint update. Hold all other parameters fixed. For a direction u, step size $\eta > 0$, and locally Lipschitz loss gradients,

$$\mathcal{E}_B(\vartheta - \eta u) - \mathcal{E}_B(\vartheta) = -\eta \langle g_B, u \rangle + O(\eta^2). \tag{2}$$

A positive inner product gives first-order descent for that band, a negative one gives an increase, and a zero inner product requires higher-order analysis. All normalized quantities below require nonzero denominators.

3.2 Why does output orthogonality not ensure gradient independence?

The band residuals are orthogonal, but their gradients pass through shared parameters. Let $J_B = \partial \mathbf{r}_B / \partial \vartheta$ be the residual Jacobian. The chain rule gives

$$g_B = J_B^\top \mathbf{r}_B, \qquad \langle g_L, g_H \rangle = \mathbf{r}_L^\top \underbrace{J_L J_H^\top}_{K_{LH}} \mathbf{r}_H. \tag{3}$$

Here K_{LH} is the cross-band Gram block, which describes how shared parameters couple the two residual bands. Gradient orthogonality depends on the shared Jacobian and current residuals, not merely on output orthogonality. Let $J = \partial \mathbf{r} / \partial \vartheta$ and define the Gram matrix $K = JJ^\top$. Symmetry and complementarity give

$$K_{LH} = \Pi_L K \Pi_H = 0 \quad \Longleftrightarrow \quad K \Pi_L = \Pi_L K. \tag{4}$$

With the Jacobian fixed, this condition means that K preserves each frequency band, making the gradient inner product zero for every residual pair. A nonzero cross block permits interaction; Eq. (3) determines its sign for the current residuals.

3.3 When does total error decrease while high-frequency error increases?

A negative inner product between the band gradients makes a sufficiently small low-frequency descent step increase high-frequency error to first order. Whether a joint update also increases high-frequency error depends on the gradient magnitudes. Define the low-to-high norm ratio and cosine similarity

$$R = \frac{\|g_L\|_2}{\|g_H\|_2}, \qquad \rho = \frac{\langle g_L, g_H \rangle}{\|g_L\|_2 \|g_H\|_2}. \tag{5}$$

Proposition 1 (Condition for increasing high-frequency error)**.** *For the joint update* $\vartheta^+ = \vartheta - \eta(g_L + g_H)$*,*

$$\mathcal{E}_H(\vartheta^+) - \mathcal{E}_H(\vartheta) = -\eta \|g_H\|_2^2 M_H + O(\eta^2), \qquad M_H = 1 + R\rho. \tag{6}$$

The unit term in M_H is high-frequency self-descent; $R\rho$ is the relative effect of the low-frequency gradient. High-frequency error increases to first order precisely when $M_H < 0$, equivalently $\rho < -1/R$. This requires $R > 1$: the opposing low-frequency effect must exceed high-frequency self-descent. Total error still decreases to first order, with low-frequency improvement compensating for high-frequency deterioration.

Weighting the losses changes the relative size of their gradient contributions.

Corollary 1 (Weights for simultaneous descent)**.** *For fixed positive weights* w_L, w_H*, let* $d^w = w_L g_L + w_H g_H$ *and* $q = w_H / w_L$*. Then*

$$\begin{aligned} \langle g_H, d^w \rangle &= \|g_H\|_2^2 M_H^w, & M_H^w &= w_H + w_L R\rho, \\ \langle g_L, d^w \rangle &= w_L \|g_H\|_2^2 (R^2 + qR\rho). \end{aligned} \tag{7}$$

High-frequency descent requires $q > -R\rho$. Requiring low-frequency descent as well gives, for $-1 < \rho < 0$,

$$\boxed{-R\rho < q < -\frac{R}{\rho}}. \tag{8}$$

The bounds protect high- and low-frequency descent, respectively. If $\rho \geq 0$, every positive ratio gives strict first-order descent in both bands; if $\rho = -1$, none does.

For comparison of unweighted and weighted updates, the experiments use the normalized contributions

$$\phi_H = \frac{\langle g_H, g_L + g_H \rangle}{\|g_L + g_H\|_2^2}, \qquad \phi_H^w = \frac{\langle g_H, d^w \rangle}{\|d^w\|_2^2}. \tag{9}$$

Their signs agree with M_H and M_H^w, respectively. A common positive rescaling of the weights preserves the sign but changes ϕ_H^w's magnitude. We therefore use the sign to assess local descent and the magnitude to describe the measured response. The analysis fixes the minibatch, timestep, and parameter block; Appendix B gives the proofs and finite-step remainder bounds.

4 From the Gradient Condition to SpecDiffusion

4.1 Frequency-Aware Loss

Frequency-Aware Loss (FAL) splits the prediction residual r_θ into low- and high-frequency terms using complementary orthogonal projections P_L, P_H, then weights them by the current noise level:

$$\mathcal{L}_{\text{FAL}} = \frac{1}{2} \mathbb{E}\left[w_L(t) \|P_L r_\theta\|_2^2 + w_H(t) \|P_H r_\theta\|_2^2 \right]. \tag{10}$$

The expectation covers data, noise, and diffusion time. At a fixed minibatch and timestep, its gradient is the weighted direction d^w.

The theory gives weight ratios for simultaneous descent. We choose a linear schedule that emphasizes low frequencies for global structure at high noise and high frequencies for detail at low noise:

$$w_L(t) = 1 + \bar{t}, \qquad w_H(t) = 2 - \bar{t}, \qquad \bar{t} = \frac{t}{T-1}. \tag{11}$$

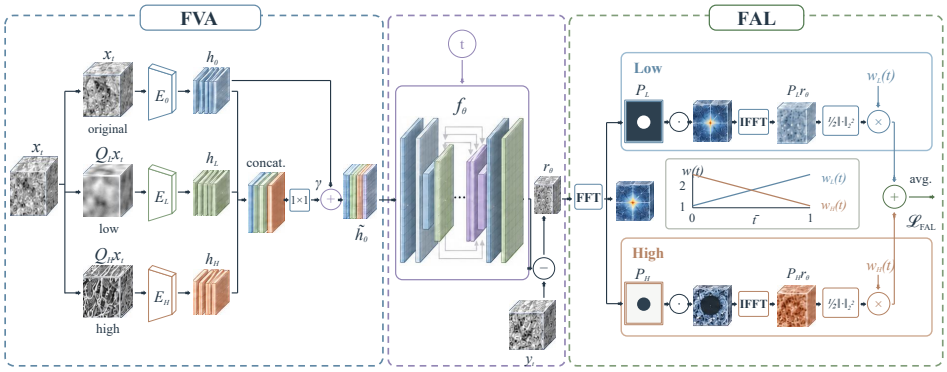

Figure 2: SpecDiffusion training. FVA fuses local input views through a gated residual branch. FAL splits the prediction residual into complementary bands, emphasizing low frequencies at high noise and high frequencies at low noise. The prediction target and reverse sampler remain unchanged.

Here T is the number of diffusion timesteps; t indexes noise level, not training progress. The weights follow directly from t, with $q(t) = w_H(t)/w_L(t) \in [1/2, 2]$. For fixed gradients with norm ratio R, cosine similarity ρ, and $M_H < 0$, the weighted update gives strict first-order high-frequency descent when

$$1 < -R\rho < q(t) \leq 2. \tag{12}$$

4.2 Spectral Partition and Loss Implementation

FAL uses a fixed Fourier partition of the prediction residual. With the axial Nyquist frequency $\omega_N = 0.5$ cycles/pixel, normalize a frequency coordinate ω by

$$\nu(\omega) = \frac{\|\omega\|_2}{\omega_N}. \tag{13}$$

We set $c = 0.35$: frequencies with $\nu(\omega) \leq c$ are low, and the rest are high. This cutoff is 0.175 cycles/pixel; the low-frequency disk covers approximately 9.6% of a square Fourier grid. The same partition applies across resolutions, datasets, and backbones.

Different band weights change both their relative importance and the timestep's overall loss scale. To separate these effects, let d be the per-sample residual dimension and $\pi_B = \mathrm{tr}(P_B)/d$ the fraction in band B, where tr denotes matrix trace. Define

$$\bar{w}(t) = \pi_L w_L(t) + \pi_H w_H(t), \qquad \widetilde{\mathcal{W}}_t = \frac{w_L(t)P_L + w_H(t)P_H}{\bar{w}(t)}. \tag{14}$$

Here $\bar{w}(t)$ is the timestep scale and $\widetilde{\mathcal{W}}_t$ the band weighting after removing that scale. Scaling both weights together changes only the former, preserving their ratio and the high-frequency descent condition. With this partition, $\bar{w}(t) \simeq 1.904 - 0.808\bar{t}$; the experiments evaluate the combined effect of timestep and relative band weighting.

The L1 component study keeps the same partition and weights, replacing squared errors with absolute errors:

$$\mathcal{L}^{(1)}_{\mathrm{FAL}} = \mathbb{E}[w_L(t)\|P_L r_\theta\|_1 + w_H(t)\|P_H r_\theta\|_1]\,. \tag{15}$$

Orthogonality exactly decomposes the squared loss. The L1 form directly defines two band losses rather than decomposing the original L1 loss. At differentiable points, gradient inner products determine the first-order band-error changes in both cases.

4.3 Frequency View Adapter

FAL weights prediction errors; the Frequency View Adapter (FVA) provides local low- and high-frequency views at the input. For a noisy input x_t, a fixed channel-wise 5×5 binomial blur $\mathcal{B}_{\mathrm{blur}}$

with reflection padding gives

$$Q_L x_t = \mathcal{B}_{\text{blur}}(x_t), \qquad Q_H x_t = x_t - \mathcal{B}_{\text{blur}}(x_t). \tag{16}$$

The views are embedded separately and fused with the original representation through a zero-initialized residual branch. Lightweight convolutions or input projections for token models keep the prediction head and sampler unchanged. FVA constructs local input views; FAL applies a global Fourier projection to the prediction residual. Component experiments measure FVA's additional effect beyond loss weighting.

5 Experiments I: Testing the Interaction Hypothesis

5.1 Diagnostic protocol

We examine three empirical questions suggested by the analysis: whether the current band gradients conflict, whether the joint direction increases high-frequency error to first order, and where a low-frequency test update changes output error. The diagnostic model is a pixel-space UNet-DDPM (Ho et al., 2020).

At each saved checkpoint, we hold the model fixed and measure the shared input block `init_conv.weight`. For each timestep, we compute g_L and g_H on each fixed diagnostic mini-batch, derive R, ρ, and ϕ_H, then take the median of each derived statistic across minibatches. We evaluate $t \in \{100, 250, 400, 550, 700, 850\}$ under L2 and L1 band losses with the same $c = 0.35$ partition. The $t = 250$ summary uses the 14 post-warmup checkpoints shown for each loss in Figure 3.

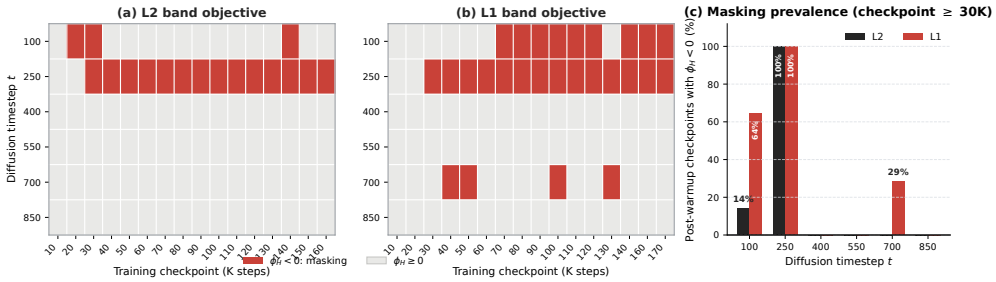

Figure 3: Testing $M_H < 0$ in a UNet-DDPM's shared input block. Cells show median ϕ_H over diagnostic minibatches at each checkpoint and diffusion timestep: red is negative, gray is nonnegative. Panels (a–b) use L2 and L1 band losses; (c) gives the fraction of post-warmup checkpoints with negative contributions. The $t = 250$ row is negative at all 14 post-warmup checkpoints for both losses.

At $t = 250$, across 14 post-warmup checkpoints, the median R is 28.91 for L2 and 29.25 for L1; the corresponding median ϕ_H values are -0.0046 and -0.0059. Under both losses, $\phi_H < 0$ at all 14 checkpoints.

5.2 From parameter-space sign to frequency-resolved effect

Figure 3 directly tests Proposition 1 in the measured block. For a gradient pair with $R \simeq 29$, high-frequency error increases to first order when ρ falls below approximately -0.034. The full checkpoint–timestep map shows how the sign varies with noise level.

After a normalized low-frequency test update, destructive-transfer rates above the cutoff are 46.5%, 55.3%, and 51.0% at early, middle, and late stages, compared with 3.1%, 14.0%, and 11.8% below the cutoff. The high–low gap remains 39.2–43.4 percentage points. The joint-gradient test in Figure 3 shows when low-frequency opposition overcomes high-frequency self-descent. The test updates in Figure 4 show that low-frequency descent increases output error more often above the cutoff across all three stages.

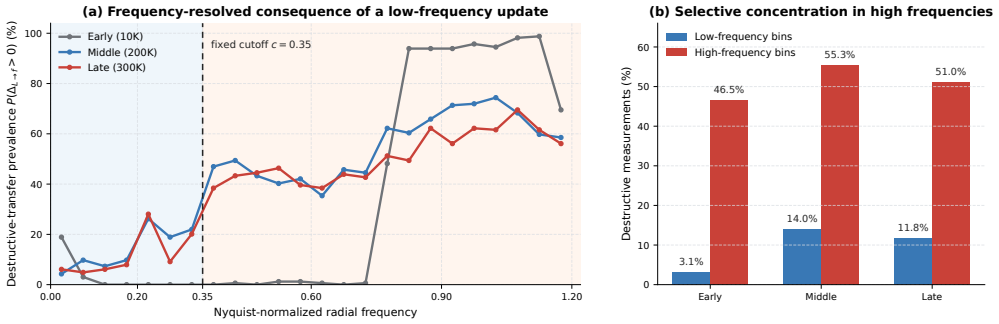

Figure 4: Error increases after a normalized low-frequency test update in the same UNet-DDPM. Destructive transfer means an increase in squared residual error. Left: the fraction of measurements with increased error in each radial bin at early, middle, and late training stages. Right: the measurements pooled below and above $c = 0.35$. Error increases are more frequent in the high-frequency bins at all three stages.

6 Experiments II: Evaluating SpecDiffusion

6.1 Matched evaluation protocol

We use CIFAR-100 (Krizhevsky, 2009), CelebA (Liu et al., 2015), FFHQ (Karras et al., 2019), LSUN-Church (Yu et al., 2015), and ImageNet (Deng et al., 2009). Backbones include UNet-DDPM, ADM (Dhariwal & Nichol, 2021), LDM (Rombach et al., 2022), Pixel-DiT (Peebles & Xie, 2023), and Lightning-DiT (Yao et al., 2025). Every baseline/intervention pair uses the same backbone, data split, optimizer, training budget, sampler, and FID implementation (Heusel et al., 2017). Configurations marked "Full" add FVA to the input representation; Lightning-DiT uses FAL alone. UNet-DDPM uses 1000-step DDPM and Pixel-DiT uses 200-step DDIM; the remaining families use 250-step DDIM. All matched FIDs use 50K generated samples and the same fixed $c = 0.35$ mask.

6.2 Gradient-contribution response to FAL

We evaluate how the local gradient geometry differs between baseline and FAL-trained CelebA models. Each model is evaluated using the same parameter block, minibatch protocol, and partition, and its own objective gradient:

$$\begin{aligned} d^{\mathrm{Base}} &= g_L^{\mathrm{Base}} + g_H^{\mathrm{Base}}, \\ d^{\mathrm{FAL}} &= w_L g_L^{\mathrm{FAL}} + w_H g_H^{\mathrm{FAL}}, \\ \phi_H^{\mathrm{Base}} &= \frac{\langle g_H^{\mathrm{Base}}, d^{\mathrm{Base}} \rangle}{\|d^{\mathrm{Base}}\|_2^2}, \qquad \phi_H^{\mathrm{FAL}} = \frac{\langle g_H^{\mathrm{FAL}}, d^{\mathrm{FAL}} \rangle}{\|d^{\mathrm{FAL}}\|_2^2}. \end{aligned} \tag{17}$$

At every diffusion timestep, we then compute

$$\Delta\phi_H(t) = \phi_H^{\mathrm{FAL}}(t) - \phi_H^{\mathrm{Base}}(t), \tag{18}$$

where the weights are evaluated at the same t. The signs of ϕ_H^{Base} and ϕ_H^{FAL} match their respective margins by Eq.(9). We compare the low- and high-noise ranges on either side of the weight crossing in Eq.(11). Both learned gradients and loss weights differ between models, so $\Delta\phi_H$ describes their contributions after training.

In the low-noise half ($t = 1$–499), the mean shift is $+5.43 \times 10^{-4}$ and 73.9% of timesteps move upward. All 27 timesteps with negative baseline contribution improve, 22 cross zero, and only five retain a negative weighted margin. In the high-noise half ($t = 500$–999), the mean shift is -1.19×10^{-2} and 98.8% move downward. The FAL-trained model therefore has fewer low-noise timesteps at which its joint update would increase high-frequency error to first order.

Table 1: Evaluation coverage and matched sampling protocol. All comparisons use 50K generated samples. UNet-DDPM, Pixel-DiT, and ADM use pixel space; LDM uses autoencoder latents and Lightning-DiT uses VA-VAE latents. Multiple datasets in one row share the same model-family protocol.

Backbone	Datasets	Resolution	Steps	Sampler
UNet-DDPM	FFHQ, CelebA, LSUN-Church	128^2	1000	DDPM
Pixel-DiT	CIFAR-100, CelebA	32^2, 64^2	200	DDIM
ADM	CIFAR-100, CelebA, FFHQ	32^2, 64^2	250	DDIM
LDM	CelebA, FFHQ, LSUN-Church	128^2, 256^2	250	DDIM
Lightning-DiT	ImageNet	256^2	250	DDIM

Table 2: Component-addition ablation on FFHQ 256^2 over 400K training steps. FID is mean $\pm$ standard deviation over three independent runs; Step is the checkpoint attaining the reported value.

Configuration	L1 Best FID	L1 Step	L2 Best FID	L2 Step
Baseline	14.04 ± 0.12	260K	18.46 ± 0.54	250K
FAL	12.47 ± 0.45	140K	$\mathbf{15.75 \pm 0.26}$	230K
FAL+FVA	$\mathbf{12.09 \pm 0.33}$	90K	15.89 ± 0.81	170K

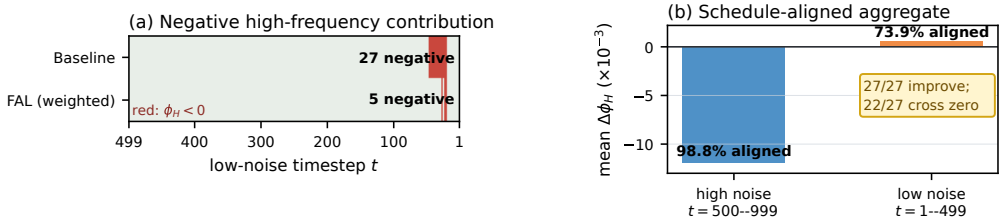

Figure 5: Gradient contributions after baseline and FAL training, using Eq.(17). Left: at low noise, negative high-frequency contributions (red) occur at 27 timesteps for the baseline and 5 for FAL. Right: mean $\Delta\phi_H$ is positive at low noise and negative at high noise; percentages give the fractions of timesteps with these respective shift signs.

6.3 Objective and Representation Components

Table 2 separates the effects of FAL and FVA. FAL alone reduces mean FID from 14.04 to 12.47 for L1 and from 18.46 to 15.75 for L2. FVA yields the best L1 mean (12.09), whereas FAL alone yields the best L2 mean (15.75 versus 15.89 for the combined configuration). Thus, loss reweighting alone lowers mean FID in both norm settings; the adapter provides a further mean FID reduction in the L1 setting.

6.4 Matched Generative Results

Table 3 compares the intervention across pixel-space and latent-space denoisers. The pixel-space rows cover convolutional and Transformer architectures. FID decreases by 1.34–2.17 for UNet-DDPM and on all three ADM datasets. Pixel-DiT reduces FID by 3.16 on CIFAR-100 and 1.01 on CelebA.

LDM and Lightning-DiT apply the same operator to spatial latent tensors. These rows test band-wise reweighting of spatial latent features; the pixel-space diagnostics above measure interaction between image frequencies. LDM reduces FID by 0.44–1.31, and the FAL-only Lightning-DiT integration reduces FID from 2.163 to 2.028. All 12 evaluated configurations yield lower FID, supporting the intervention beyond the diagnostic backbone in both pixel and latent spaces.

Table 3: Complete matched-baseline FID comparison under the common $c = 0.35$ cutoff; lower is better. Except for Lightning-DiT, which is reported from a single run, results are mean $\pm$ standard deviation over three independent seeds.

Backbone	Space	Dataset	Realization	Baseline	Ours
UNet-DDPM	Pixel	FFHQ128	Full	13.24 ± 0.18	$\mathbf{11.07 \pm 0.16}$
UNet-DDPM	Pixel	CelebA128	Full	11.38 ± 0.14	$\mathbf{9.88 \pm 0.12}$
UNet-DDPM	Pixel	LSUN-Church128	Full	14.77 ± 0.20	$\mathbf{13.43 \pm 0.18}$
LDM	Latent	CelebA128	Full	19.54 ± 0.22	$\mathbf{19.10 \pm 0.20}$
LDM	Latent	LSUN-Church256	Full	10.45 ± 0.12	$\mathbf{9.14 \pm 0.11}$
LDM	Latent	FFHQ256	Full	15.23 ± 0.19	$\mathbf{14.39 \pm 0.03}$
ADM	Pixel	CIFAR-100	Full	6.295 ± 0.07	$\mathbf{5.828 \pm 0.06}$
ADM	Pixel	CelebA64	Full	10.253 ± 0.11	$\mathbf{8.042 \pm 0.09}$
ADM	Pixel	FFHQ64	Full	7.830 ± 0.09	$\mathbf{6.801 \pm 0.08}$
Pixel-DiT	Pixel	CIFAR-100	Full	16.94 ± 0.21	$\mathbf{13.78 \pm 0.19}$
Pixel-DiT	Pixel	CelebA64	Full	14.31 ± 0.19	$\mathbf{13.30 \pm 0.17}$
Lightning-DiT	Latent	ImageNet256	FAL	2.163	**2.028**

6.5 Weight schedule evidence

The supplementary CIFAR-100 schedule study evaluates linear, cosine, quadratic, and step schedules with a common 50-step DDIM sampler and 10K generated samples, separately from the 50K-sample main comparisons. The three smooth schedules show relatively small differences in final FID and all provide noise-dependent band weighting, motivating the simple linear schedule used here.

7 Conclusion and Limitations

This work offers a gradient-interaction view of frequency-band learning in diffusion models with shared parameters. Guided by this analysis, we propose Frequency-Aware Loss (FAL), which improves generation quality across the evaluated pixel- and latent-space denoisers. The current method uses a fixed two-band partition and a simple bounded schedule. Future work could explore finer-grained or continuous spectral modeling, as well as band interactions across network layers, under actual optimizer updates, and with adaptive weights to guide frequency-aware training.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, derivation checking, document-format conversion, and visual refinement of Figures 1 and 2. The authors reviewed the resulting text and figures against the manuscript and experiment records and remain responsible for all scientific claims, derivations, experiments, and reported results.

REFERENCES

Zhennan Chen, Junwei Zhu, Xu Chen, Jiangning Zhang, Xiaobin Hu, Hanzhen Zhao, Chengjie Wang, Jian Yang, and Ying Tai. DiP: Taming diffusion models in pixel space. *arXiv preprint arXiv:2511.18822*, 2025.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: Conditioning method for denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Seungho Choi, Jeahun Sung, and Jihyong Oh. FRAMER: Frequency-aligned self-distillation with adaptive modulation leveraging diffusion priors for real-world image super-resolution. *arXiv preprint arXiv:2512.01390*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Yunhui Gao and Liangcai Cao. A complex constrained total variation image denoising algorithm with application to phase retrieval, 2021. URL `https://arxiv.org/abs/2109.05496`.

Hyojun Go, JinYoung Kim, Yunsung Lee, Seunghyun Lee, Shinhyeok Oh, Hyeongdon Moon, and Seungtaek Choi. Addressing negative transfer in diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-SNR weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Xingchang Huang, Corentin Salaün, Cristina Vasconcelos, Christian Theobalt, Cengiz Öztireli, and Gurprit Singh. Blue noise for diffusion models. *arXiv preprint arXiv:2402.04930*, 2024.

Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems*, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

Tao Luo, Zheng Ma, Zhi-Qin John Xu, and Yaoyu Zhang. Theory of the frequency principle for general deep neural networks. *CSIAM Transactions on Applied Mathematics*, 2(3):484–507, 2021.

Qianli Ma, Xuefei Ning, Dongrui Liu, Li Niu, and Linfeng Zhang. Decouple-then-merge: Finetune diffusion models as multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Zehong Ma, Longhui Wei, Shuai Wang, Shiliang Zhang, and Qi Tian. DeCo: Frequency-decoupled pixel diffusion for end-to-end image generation. *arXiv preprint arXiv:2511.19365*, 2025b.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

David Mueller, Mark Dredze, and Nicholas Andrews. Can optimization trajectories explain multi-task transfer? *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=QQE5j2OsLW`.

Mang Ning, Mingxiao Li, Jianlin Su, Haozhe Jia, Lanmiao Liu, Martin Benes, Wenshuo Chen, Albert Ali Salah, and Itir Onal Ertugrul. DCTdiff: Intriguing properties of image generative modeling in the DCT space. In *International Conference on Machine Learning*, pp. 46498–46524, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Katja Schwarz, Yiyi Liao, and Andreas Geiger. On the frequency bias of generative models. In *Advances in Neural Information Processing Systems*, 2021.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, 2018.

Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3):391–412, 2003.

Binxu Wang and Cengiz Pehlevan. An analytical theory of spectral bias in the learning dynamics of diffusion models. *arXiv preprint arXiv:2503.03206*, 2025.

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767, 2020.

Chao Yang, Boqian Zhang, Jinghao Xu, and Guang Jiang. HDW-SR: High-frequency guided diffusion model based on wavelet decomposition for image super-resolution. *arXiv preprint arXiv:2511.13175*, 2025.

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

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, 2020.

Yu Zeng, Charles Ochoa, Mingyuan Zhou, Vishal M. Patel, Vitor Guizilini, and Rowan McAllister. NeuralRemaster: Phase-preserving diffusion for structure-aligned generation. *arXiv preprint arXiv:2512.05106*, 2025.

Juyan Zhang, Rhys Newbury, Xinyang Zhang, Tin Tran, Dana Kulic, and Michael Burke. Heteroscedasticity of denoising score matching with generalised smooth noise. *arXiv preprint arXiv:2508.01597v2*, 2026. URL `https://arxiv.org/abs/2508.01597v2`.

A Overview and Notation

The main paper studies how a shared denoiser's update changes low- and high-frequency errors. This supplement develops the local derivations and examines what minibatch measurements imply for population risk. It then documents the gradient diagnostics, implementation, and matched evaluation protocols. The experimental sections use the main paper's spectral partition and baseline/intervention pairs.

For the local analysis, fix a diffusion timestep t, a minibatch $\mathcal{B} = \{(x_t^{(i)}, y_t^{(i)})\}_{i=1}^m$ of m input–target pairs, and a parameter block ϑ within the full parameters θ. The choice $\vartheta = \theta$ covers the whole model. Let $f_\theta(x_t, t)$ be the denoiser prediction and y_t its prediction target. Their difference is the residual $r_\theta(x_t, t) = f_\theta(x_t, t) - y_t$. The complementary orthogonal projections P_L and P_H retain the low- and high-frequency parts of that residual. For $B \in \{L, H\}$, stack the projected residuals:

$$\mathbf{r}_B = \frac{1}{\sqrt{m}} \begin{bmatrix} P_B r_\theta(x_t^{(1)}, t) \\ \vdots \\ P_B r_\theta(x_t^{(m)}, t) \end{bmatrix}. \tag{A.1}$$

The factor $m^{-1/2}$ makes the squared norm the minibatch-average squared error. Define the band error $\mathcal{E}_B$, its gradient g_B with respect to the selected block, and its Jacobian J_B:

$$\mathcal{E}_B = \frac{1}{2}\|\mathbf{r}_B\|_2^2, \qquad g_B = \nabla_\vartheta \mathcal{E}_B, \qquad J_B = \frac{\partial \mathbf{r}_B}{\partial \vartheta}. \tag{A.2}$$

All three quantities refer to the same block, minibatch, and timestep. Later ratios require their denominators to be nonzero. When a gradient vanishes, the unnormalized change formula in Eq.(B.5) still applies. Section C keeps the model and timestep fixed while allowing the minibatch to vary; it introduces a separate symbol for each population gradient.

B Complete Derivations

Orthogonal residual bands can interact through shared parameters; the cross-band Gram block identifies when that interaction vanishes. A finite-step expansion distinguishes low-frequency test updates from joint updates. Applying the same expansion to weighted gradients gives the high-frequency descent condition and the interval where both bands improve.

B.1 Exact loss decomposition and shared-parameter coupling

The projections satisfy $P_L + P_H = I$ and $P_L P_H = 0$, where I is the identity on one residual. Orthogonality then splits the squared error on the fixed minibatch exactly:

$$\mathcal{E} = \mathcal{E}_L + \mathcal{E}_H, \qquad \nabla_\vartheta \mathcal{E} = g_L + g_H. \tag{B.1}$$

The first equality concerns error in output space. Both gradients in the second equality pass through the same selected parameters. By the chain rule, $g_B = J_B^\top \mathbf{r}_B$, so their inner product is

$$\langle g_L, g_H \rangle = \mathbf{r}_L^\top \underbrace{J_L J_H^\top}_{K_{LH}} \mathbf{r}_H. \tag{B.2}$$

Here $K_{LH} = J_L J_H^\top$ is the cross-band block of the Jacobian Gram matrix. It describes how an update driven by one residual band changes the other. If $K_{LH} = 0$, no residual pair has such a first-order interaction through this parameter block.

The condition $K_{LH} = 0$ can also be expressed directly in terms of the Jacobian. Let $\mathbf{r}$ be the unprojected batch residual with the same $m^{-1/2}$ normalization, let $J = \partial \mathbf{r}/\partial \vartheta$ be its Jacobian, and set $K = JJ^\top$. As in the main paper, $\Pi_B = \mathrm{diag}(P_B, \ldots, P_B)$ applies the same projection to each of the m samples, and $K_{LH} = \Pi_L K \Pi_H$. Because K is symmetric and $\Pi_L + \Pi_H = I$, we have

$$K_{LH} = 0 \quad \Longleftrightarrow \quad K\Pi_L = \Pi_L K. \tag{B.3}$$

For the forward direction, transpose the zero cross block and use $I = \Pi_L + \Pi_H$ to obtain $K\Pi_L = \Pi_L K \Pi_L = \Pi_L K$. For the reverse direction, the commutation relation gives $\Pi_L K \Pi_H = K \Pi_L \Pi_H = 0$. The operator condition removes interaction for every residual pair. A particular pair can still have zero gradient inner product when K_{LH} is nonzero, as Eq.(B.2) shows.

B.2 FINITE-STEP CHANGE

Assumption 1 (Local smoothness). *For $B \in \{L, H\}$, the error $\mathcal{E}_B$ is differentiable near the measured parameter value. Its gradient is locally Lipschitz with constant $L_B \geq 0$.*

For a fixed direction u and sufficiently small step size $\eta > 0$, local smoothness gives

$$|\mathcal{E}_B(\vartheta - \eta u) - \mathcal{E}_B(\vartheta) + \eta\langle g_B, u\rangle| \leq \frac{L_B\eta^2}{2}\|u\|_2^2. \tag{B.4}$$

Take positive band weights w_L, w_H fixed during the update, form the weighted gradient $d^w = w_L g_L + w_H g_H$, and update $\vartheta^+ = \vartheta - \eta d^w$ with step size $\eta > 0$. Putting $u = d^w$ into Eq.(B.4) gives

$$\mathcal{E}_B(\vartheta^+) - \mathcal{E}_B(\vartheta) = -\eta\langle g_B, d^w\rangle + \xi_B, \qquad |\xi_B| \leq \frac{L_B\eta^2}{2}\|d^w\|_2^2. \tag{B.5}$$

The symbol ξ_B is the difference between the exact change and its first-order term. Define the band directional quantity $a_B^w = \langle g_B, d^w\rangle$. When $a_B^w \neq 0$ and $L_B > 0$, the exact finite change has the sign of $-a_B^w$ if the step stays in the smooth neighborhood and

$$0 < \eta < \frac{2|a_B^w|}{L_B\|d^w\|_2^2}. \tag{B.6}$$

Under this bound the remainder is smaller than $\eta|a_B^w|$. If $L_B = 0$, the remainder vanishes locally. When $a_B^w = 0$, the first-order term alone does not determine the change.

B.3 A TEST UPDATE ALONG THE LOW-FREQUENCY GRADIENT

Assume $g_L \neq 0$ and $\mathcal{E}_H > 0$, and normalize the low-frequency gradient as $\widehat{g}_L = g_L/\|g_L\|_2$. A test step along its negative direction changes the high-frequency error by the relative amount

$$\Delta_{L\to H} = \frac{\mathcal{E}_H(\vartheta - \eta\widehat{g}_L) - \mathcal{E}_H(\vartheta)}{\mathcal{E}_H(\vartheta)}. \tag{B.7}$$

Theorem 1 (Low-to-high error change). *Under local smoothness,*

$$\Delta_{L\to H} = -\eta\frac{\langle g_H, g_L\rangle}{\mathcal{E}_H\|g_L\|_2} + \xi_{L\to H}, \qquad |\xi_{L\to H}| \leq \frac{L_H\eta^2}{2\mathcal{E}_H}. \tag{B.8}$$

Here $\xi_{L\to H}$ is the remainder after dividing by $\mathcal{E}_H$. If both gradients are nonzero, their cosine similarity $\rho = \langle g_L, g_H\rangle/(\|g_L\|_2\|g_H\|_2)$ makes the direction explicit:

$$\Delta_{L\to H} = -\eta\frac{\|g_H\|_2}{\mathcal{E}_H}\rho + O(\eta^2). \tag{B.9}$$

The notation $O(\eta^2)$ denotes a remainder bounded by a constant times η^2 for sufficiently small η.

Proof. Apply Eq.(B.4) with $B = H$ and $u = g_L/\|g_L\|_2$, then divide by $\mathcal{E}_H$. □

Reversing the two bands uses the same cosine but a different size factor. When both gradients are nonzero and $\mathcal{E}_L > 0$, the analogous relative change is

$$\Delta_{H\to L} = -\eta\frac{\|g_L\|_2}{\mathcal{E}_L}\rho + O(\eta^2). \tag{B.10}$$

The two relative changes need not have the same magnitude.

B.4 WHEN A JOINT UPDATE INCREASES HIGH-FREQUENCY ERROR

The test step above uses only the low-frequency gradient. A joint step also contains the high-frequency gradient's own descent effect. To compare the two, define their norm ratio $R = \|g_L\|_2/\|g_H\|_2$ for nonzero band gradients.

Proposition 2 (High-frequency change in a selected block)**.** *For the unweighted joint direction* $g_L + g_H$*,*

$$\mathcal{E}_H(\vartheta - \eta(g_L + g_H)) - \mathcal{E}_H(\vartheta) = -\eta\|g_H\|_2^2 M_H + O(\eta^2), \qquad M_H = 1 + R\rho. \tag{B.11}$$

The first-order term increases the high-frequency error when

$$M_H < 0 \quad \Longleftrightarrow \quad \rho < -\frac{1}{R}. \tag{B.12}$$

Proof. Set $B = H$ and $w_L = w_H = 1$ in Eq.(B.5), and use $\langle g_H, g_L\rangle = R\rho\|g_H\|_2^2$. □

The experiments also report the normalized contribution

$$\phi_H = \frac{\langle g_H, g_L + g_H\rangle}{\|g_L + g_H\|_2^2} = \frac{\|g_H\|_2^2 M_H}{\|g_L + g_H\|_2^2}. \tag{B.13}$$

Its denominator is positive whenever $g_L + g_H \neq 0$. Its sign then agrees with M_H, although its magnitude uses a different normalization.

B.5 How band weights change the high-frequency condition

Keep the same gradients and change their positive weights in $d^w = w_L g_L + w_H g_H$. Let $q = w_H/w_L$ be the high-to-low weight ratio.

Corollary 2 (Weighted high-frequency change)**.** *For a step along* $-d^w$*,*

$$\mathcal{E}_H(\vartheta - \eta d^w) - \mathcal{E}_H(\vartheta) = -\eta\|g_H\|_2^2 M_H^w + O(\eta^2), \tag{B.14}$$

where the weighted margin is

$$M_H^w = w_H + w_L R\rho. \tag{B.15}$$

The high-frequency error strictly decreases to first order when

$$q > -R\rho. \tag{B.16}$$

Proof. The two band directional quantities are

$$\begin{aligned} \langle g_H, d^w\rangle &= w_L\|g_H\|_2^2(q + R\rho), \\ \langle g_L, d^w\rangle &= w_L\|g_H\|_2^2(R^2 + qR\rho). \end{aligned} \tag{B.17}$$

Substitute the first identity into Eq.(B.5) with $B = H$. □

For a nonzero weighted direction, the normalized contribution is

$$\phi_H^w = \frac{\langle g_H, d^w\rangle}{\|d^w\|_2^2} = \frac{\|g_H\|_2^2 M_H^w}{\|w_L g_L + w_H g_H\|_2^2}. \tag{B.18}$$

Its sign agrees with M_H^w.

B.6 Weights that make both bands decrease

Strict first-order descent in both bands requires both inner products in Eq.(B.17) to be positive. For $-1 < \rho < 0$, these two conditions give

$$-R\rho < q < -\frac{R}{\rho}. \tag{B.19}$$

The interval is nonempty. At $\rho = -1$, its endpoints coincide at $q = R$, where $d^w = 0$, so strict simultaneous descent is impossible. At $\rho \geq 0$, every positive ratio makes both leading terms positive.

The FAL schedule uses $q(t) \in [1/2, 2]$. For a fixed gradient pair whose unweighted update increases high-frequency error to first order, the schedule restores strict first-order high-frequency descent when $1 < -R\rho < q(t) \leq 2$. If $-R\rho = 2$, the maximum scheduled ratio gives a zero high-frequency directional term; if $-R\rho > 2$, no scheduled ratio restores descent for that pair. During training, the residuals and Jacobians change, and so do R and ρ.

B.7 RATIO OF FIRST-ORDER DECREASES

For a fixed gradient pair in the simultaneous-descent region, dividing the two quantities in Eq.(B.17) gives the ratio of low- to high-frequency first-order error decreases:

$$\Gamma(q) = \frac{R^2 + q\rho R}{q + \rho R}. \tag{B.20}$$

For $|\rho| < 1$, differentiation yields

$$\frac{d\Gamma}{dq} = -\frac{R^2(1-\rho^2)}{(q+\rho R)^2} < 0. \tag{B.21}$$

Increasing q therefore reduces the low-frequency decrease relative to the high-frequency decrease within this region.

The normalized contribution ϕ_H^w need not follow this ordering. Multiplying both weights by the same positive factor divides ϕ_H^w by that factor while preserving its sign. Even with fixed gradients, raising q can lower ϕ_H^w: for two orthogonal unit gradients, weights $(1, 1)$ give $1/2$, whereas $(1, 2)$ give $2/5$. Both directions decrease high-frequency error. We use the sign of ϕ_H^w for the local descent test and its magnitude to describe the measured response.

Timestep scale and band weights. The ratio $q(t)$ describes the relative band weights at one timestep. To separate these weights from their common scale, define $\mathcal{W}_t = w_L(t)P_L + w_H(t)P_H$ and $\pi_B = \mathrm{tr}(P_B)/d$, where d is the residual dimension and tr denotes matrix trace. The mean weight and normalized operator are

$$\bar{w}(t) = \pi_L w_L(t) + \pi_H w_H(t), \qquad \widetilde{\mathcal{W}}_t = \frac{\mathcal{W}_t}{\bar{w}(t)}. \tag{B.22}$$

The scalar $\bar{w}(t)$ sets the timestep scale, while $\widetilde{\mathcal{W}}_t$ specifies the band allocation and satisfies $\mathrm{tr}(\widetilde{\mathcal{W}}_t)/d = 1$. A common positive rescaling of the weights changes $\bar{w}(t)$ but preserves $\widetilde{\mathcal{W}}_t$ and $q(t)$.

C MINIBATCH DIRECTIONAL CRITERIA AND POPULATION RISK

The local criterion describes high-frequency error on the minibatch used for the update. Its relation to population risk depends on the fluctuations of minibatch gradients. We separate these fluctuations from the population gradients, examine their effect in a Gaussian denoising model, and derive an estimator of the population directional term from independent minibatches.

C.1 WHAT A MINIBATCH DIRECTIONAL QUANTITY MEASURES

Fix the parameter block ϑ and diffusion timestep t as before. Let $Z = (x_t, y_t)$ denote a random input–target pair at this timestep, and let $Z_1, \ldots, Z_m$ be independent pairs with the same distribution. For $B \in \{L, H\}$, define the single-pair loss ℓ_B, population loss $\mathcal{L}_B$, and population gradient G_B by

$$\ell_B(\vartheta; Z) = \frac{1}{2}\|P_B r_\theta(x_t, t)\|_2^2, \qquad \mathcal{L}_B(\vartheta) = \mathbb{E}_Z[\ell_B(\vartheta; Z)], \qquad G_B = \nabla_\vartheta \mathcal{L}_B. \tag{C.1}$$

The expectation $\mathbb{E}_Z$ averages over both data and forward noise. The remaining model parameters are held fixed. Assume differentiation and expectation can be interchanged and that the single-pair gradients have finite second moments. Then the minibatch gradient already denoted by g_B satisfies

$$g_B = \frac{1}{m}\sum_{i=1}^{m} \nabla_\vartheta \ell_B(\vartheta; Z_i), \qquad \mathbb{E}[g_B] = G_B. \tag{C.2}$$

Take positive weights w_L, w_H that are fixed before drawing the minibatch. The update direction remains $d^w = w_L g_L + w_H g_H$. Write

$$a_H^w = \langle g_H, d^w \rangle = \|g_H\|_2^2 M_H^w. \tag{C.3}$$

Here a_H^w is the unnormalized high-frequency directional quantity; the second equality applies when the margin is defined. A negative a_H^w means a first-order increase in the current minibatch error. For $B, B' \in \{L, H\}$, introduce the scalar

$$C_{BB'} = \mathbb{E}_Z[\langle \nabla_\vartheta \ell_B - G_B,\, \nabla_\vartheta \ell_{B'} - G_{B'} \rangle]\,. \tag{C.4}$$

Both gradients inside this expectation use the same pair Z. The value C_{BB} is the mean squared gradient fluctuation; C_{LH} is the inner product of the two fluctuations averaged over that shared pair. Expanding the minibatch averages gives

$$\mathbb{E}\langle g_B, g_{B'} \rangle = \langle G_B, G_{B'} \rangle + \frac{C_{BB'}}{m}. \tag{C.5}$$

Different-sample terms factor by independence; same-sample terms retain the gradient covariance. Applying this identity to the weighted direction yields

$$\mathbb{E}[a_H^w] = \langle G_H, w_L G_L + w_H G_H \rangle + \frac{w_L C_{HL} + w_H C_{HH}}{m}. \tag{C.6}$$

The first term determines the expected first-order change in population high-frequency loss, with coefficient $-\eta$ in the update expansion: $\mathbb{E}\langle G_H, d^w \rangle = \langle G_H, w_L G_L + w_H G_H \rangle$. The second term appears because the minibatch loss and update direction use the same samples. It can be negative even though $C_{HH} \geq 0$. This identity concerns an unnormalized quantity; averaging M_H^w or ϕ_H^w does not give the same decomposition because their denominators also depend on the minibatch.

C.2 A Gaussian denoising example

A Gaussian denoising model makes this distinction explicit. In the following statement, $\delta \in (0, 1)$ bounds the probability of a nonnegative margin.

Proposition 3 (Minibatch signs and population risk)**.** *There exists a Gaussian diffusion denoising problem with one shared trainable parameter for which both population band losses have the same unique minimizer and their gradients are strictly aligned elsewhere. Within this family, for any $\delta \in (0, 1)$ one can choose a single problem whose unweighted high-frequency margin is negative at the minimizer with probability greater than $1 - \delta$ for every finite minibatch size. There also exists a problem in the same family where changing the positive weight ratio from $q = 1$ to $q = 2$, while keeping the weight sum, step size, and minibatch size fixed, reduces the stationary probability of a negative high-frequency directional quantity but increases the stationary population high-frequency excess risk.*

We prove the two statements using the same construction. Work in two orthonormal real Fourier coordinates, with the first in the low-frequency band and the second in the high-frequency band. The projections retain these coordinates separately: $P_L = \mathrm{diag}(1, 0)$ and $P_H = \mathrm{diag}(0, 1)$. They are the restrictions of the band projections to these two modes. One may embed the construction in a larger Fourier basis by setting all other data coordinates to zero and predicting their noise exactly.

At the fixed timestep, let $a > 0$ and $b > 0$ denote the signal and noise scales. The forward process and noise prediction target are

$$x_t = a x_0 + b\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_2), \qquad y_t = \varepsilon. \tag{C.7}$$

Here $\mathcal{N}(0, \Sigma)$ denotes a zero-mean Gaussian distribution with covariance Σ, and I_2 is the two-dimensional identity matrix. Choose a residual scale $s > 0$ and a correlation parameter κ satisfying $0 < \kappa < 1$ and $s < 1/(1 + \kappa)$. Define

$$S = s \begin{bmatrix} 1 & -\kappa \\ -\kappa & 1 \end{bmatrix}, \qquad \Sigma_0 = \frac{b^2}{a^2} S (I_2 - S)^{-1}, \qquad x_0 \sim \mathcal{N}(0, \Sigma_0), \tag{C.8}$$

with x_0 independent of ε. The matrix S will be the conditional covariance of the noise, while Σ_0 is the data covariance. The eigenvalues of S are $s(1 - \kappa)$ and $s(1 + \kappa)$, both strictly between zero and one, so Σ_0 is positive definite. The covariance of x_t is $a^2\Sigma_0 + b^2 I_2 = b^2 (I_2 - S)^{-1}$. Gaussian conditioning now gives

$$\mathrm{Cov}(\varepsilon \mid x_t) = I_2 - b^2 (a^2 \Sigma_0 + b^2 I_2)^{-1} = S, \tag{C.9}$$

where Cov denotes covariance. The desired residual correlation thus arises from a valid Gaussian forward process.

Let $f^*(x_t) = \mathbb{E}[\varepsilon \mid x_t]$ be the Bayes predictor, that is, the predictor minimizing population squared error. In this construction $f^*(x_t) = b^{-1}(I_2 - S)x_t$. Choose a fixed coefficient $\lambda > 1$ controlling the sensitivity of the low-frequency prediction to the shared parameter. With $\theta = \vartheta$ a single trainable scalar, set

$$f_\theta(x_t) = f^*(x_t) + \vartheta \begin{bmatrix} \lambda \\ 1 \end{bmatrix}, \qquad r_0 = f^*(x_t) - \varepsilon. \tag{C.10}$$

The timestep argument is suppressed because it is fixed. The residual r_0 is the earlier denoising residual evaluated at $\vartheta = 0$. Its coordinate $r_{0,B}$ is the scalar residual in band B. Joint Gaussianity implies that r_0 is independent of x_t and has distribution $\mathcal{N}(0, S)$. The fixed function f^* makes the optimum explicit in this theoretical model. The training updates below act only on ϑ.

Taking expectations of the two squared residuals gives

$$\mathcal{L}_L(\vartheta) = \frac{s}{2} + \frac{\lambda^2\vartheta^2}{2}, \qquad \mathcal{L}_H(\vartheta) = \frac{s}{2} + \frac{\vartheta^2}{2}. \tag{C.11}$$

Both losses have the unique minimizer $\vartheta = 0$. Their population gradients $G_L = \lambda^2\vartheta$ and $G_H = \vartheta$ have the same sign everywhere else, since $G_L G_H = \lambda^2\vartheta^2 > 0$.

C.3 A NEGATIVE MINIBATCH MARGIN AT THE OPTIMUM

For a minibatch, let $\bar{r}_{0,B} = m^{-1}\sum_{i=1}^m r_{0,B}^{(i)}$ be the average optimal residual in band B. Differentiating the empirical losses in the scalar model yields

$$g_L = \lambda^2\vartheta + \lambda\bar{r}_{0,L}, \qquad g_H = \vartheta + \bar{r}_{0,H}, \qquad \begin{bmatrix} \bar{r}_{0,L} \\ \bar{r}_{0,H} \end{bmatrix} \sim \mathcal{N}(0, S/m). \tag{C.12}$$

At $\vartheta = 0$, the population gradients vanish, but the minibatch gradients still vary with the residuals. For positive weights with ratio $q = w_H/w_L$, the directional quantity becomes

$$a_H^w = w_L\bar{r}_{0,H}(\lambda\bar{r}_{0,L} + q\bar{r}_{0,H}). \tag{C.13}$$

Since $w_L > 0$, its sign is the sign of the product of the two displayed Gaussian factors. Denote their correlation coefficient by $\varrho(q)$:

$$\varrho(q) = \frac{q - \lambda\kappa}{\sqrt{\lambda^2 + q^2 - 2\lambda\kappa q}}. \tag{C.14}$$

The symbol $\varrho(q)$ is the correlation of two random factors when minibatches are repeatedly sampled. The earlier ρ is the cosine between two gradients on one fixed minibatch.

To calculate the sign probability, write two unit-variance Gaussian variables with correlation $\varrho \in (-1, 1)$ as $U_1 = N_1$ and $U_2 = \varrho N_1 + \sqrt{1-\varrho^2}N_2$, where N_1, N_2 are independent standard normal variables. The direction of (N_1, N_2) is uniform on the circle, and the opposite-sign regions occupy a fraction $\arccos(\varrho)/\pi$ of its angle. Consequently,

$$\mathbb{P}(a_H^w < 0) = \frac{\arccos\varrho(q)}{\pi} \quad \text{at } \vartheta = 0. \tag{C.15}$$

Here $\mathbb{P}$ is probability over minibatch sampling. Nonzero gradients and directions hold with probability one, so this is also the probability that the corresponding margin is negative.

For the unweighted case $w_L = w_H = 1$, set $q = 1$. For any fixed $\lambda > 1$, letting κ tend to one gives $\varrho(1) \to -1$ and hence $\mathbb{P}(M_H < 0) \to 1$. One can keep $s = 1/4$ throughout this limit. Since s/m cancels from the correlation, a single choice of κ gives probability greater than $1 - \delta$ for every finite m, proving the first part of Proposition 3.

Larger batches still reduce the gradient amplitudes: at the optimum, $\mathbb{E}[a_H^w] = w_L(s/m)(q - \lambda\kappa)$. The sign probability stays fixed while the directional effect shrinks; in the infinite-batch limit both gradients vanish and the normalized margin is undefined. For the chosen unweighted problem and each fixed finite m, the Gaussian pair has positive-definite covariance, so the sign probability is continuous in ϑ. It remains above $1 - \delta$ in a neighborhood of the optimum, including points where $G_L G_H > 0$; that neighborhood may depend on m.

C.4 How reweighting changes population risk during continued training

The previous calculation concerns a fixed optimal parameter. To compare training outcomes, retain the same model and draw an independent fresh minibatch at every iteration. Set

$$w_L = \frac{1}{1+q}, \qquad w_H = \frac{q}{1+q}, \qquad q > 0. \tag{C.16}$$

The sum of the weights is one for all q, so the comparison changes their relative allocation without a common rescaling. Using Eq.(C.12), write the update direction as

$$d^w = \mathcal{H}(q)\vartheta + \zeta_q, \qquad \mathcal{H}(q) = \frac{\lambda^2 + q}{1+q}, \qquad \zeta_q = \frac{\lambda \bar{r}_{0,L} + q\bar{r}_{0,H}}{1+q}. \tag{C.17}$$

The scalar $\mathcal{H}(q)$ is the curvature of the weighted population loss. The zero-mean variable ζ_q is its minibatch gradient noise. Define its variance as $\sigma^2(q) = \mathrm{Var}(\zeta_q)$, where Var denotes variance. From S/m,

$$\sigma^2(q) = \frac{s}{m}\frac{\lambda^2 + q^2 - 2\lambda\kappa q}{(1+q)^2}. \tag{C.18}$$

For a fixed step size $\eta > 0$, SGD obeys the exact recursion

$$\vartheta_{k+1} = (1 - \eta\mathcal{H}(q))\vartheta_k - \eta\zeta_{q,k}. \tag{C.19}$$

Here k is the training iteration, and $\zeta_{q,k}$ denotes independent copies of ζ_q. It is independent of ϑ_k because the current minibatch is drawn independently of all preceding updates.

Assume $0 < \eta\mathcal{H}(q) < 2$, so the recursion coefficient has absolute value below one. With independent Gaussian noise at each iteration, the recursion has a unique zero-mean stationary distribution. Let ϑ_∞ be a draw from this distribution. The recursion gives its variance $V(q) = \mathrm{Var}(\vartheta_\infty)$ as

$$V(q) = \frac{\eta\sigma^2(q)}{\mathcal{H}(q)(2 - \eta\mathcal{H}(q))}. \tag{C.20}$$

Define the stationary high-frequency excess risk $\mathcal{R}_H(q)$ as the population loss above its minimum, averaged over this parameter distribution. Eq.(C.11) gives

$$\mathcal{R}_H(q) = \mathbb{E}[\mathcal{L}_H(\vartheta_\infty) - \mathcal{L}_H(0)] = \frac{V(q)}{2}. \tag{C.21}$$

We evaluate the minibatch sign under the same stationary distribution. The parameter ϑ_∞ and a fresh minibatch are independent, so $g_H = \vartheta_\infty + \bar{r}_{0,H}$ and $d^w = \mathcal{H}(q)\vartheta_\infty + \zeta_q$ form a zero-mean Gaussian pair. Their variances and covariance are

$$\mathrm{Var}(g_H) = V(q) + s/m, \tag{C.22}$$

$$\mathrm{Var}(d^w) = \mathcal{H}(q)^2 V(q) + \sigma^2(q), \tag{C.23}$$

$$\mathrm{Cov}(g_H, d^w) = \mathcal{H}(q)V(q) + \frac{s}{m}\frac{q - \lambda\kappa}{1+q}. \tag{C.24}$$

These second moments determine the correlation of the pair,

$$\varrho_\infty(q) = \frac{\mathrm{Cov}(g_H, d^w)}{\sqrt{\mathrm{Var}(g_H)\,\mathrm{Var}(d^w)}}. \tag{C.25}$$

The Gaussian sign formula used above therefore yields

$$\mathbb{P}_\infty(a_H^w < 0) = \frac{\arccos \varrho_\infty(q)}{\pi}. \tag{C.26}$$

The subscript on $\mathbb{P}_\infty$ means that probability averages over both the stationary parameter and a fresh minibatch.

Choose $\lambda = 1.1$, $\kappa = 0.99$, $s = 0.25$, and $\eta = 0.1$. At $q = 1$ and $q = 2$, the curvatures are 1.105 and 1.07, respectively, so both recursions satisfy the stability condition. Direct substitution yields

$$\mathcal{R}_H(1) = \frac{40}{835159m}, \qquad \mathcal{R}_H(2) = \frac{2135}{3645918m} > \mathcal{R}_H(1). \tag{C.27}$$

Table C.1: Closed-form values for the Gaussian construction. These values are evaluations of Eq.(C.21) and Eq.(C.26), not measurements of a trained diffusion network. Both columns use the same step size and batch size.

Quantity	$q = 1$	$q = 2$
Stationary probability $\mathbb{P}_\infty(a_H^w < 0)$	65.896%	7.465%
Relative excess risk $\mathcal{R}_H(q)/\mathcal{R}_H(1)$	1.000	12.226

The stationary covariance in Eq.(C.24) is strictly negative at $q = 1$ and strictly positive at $q = 2$. The negative-sign probability is therefore above $1/2$ in the first case and below $1/2$ in the second. These strict comparisons do not depend on rounding. Table C.1 reports their numerical values. Both risks scale as $1/m$, whereas the risk ratio and the stationary sign probabilities are independent of finite m.

This proves the second part of Proposition 3. The negative correlation between residuals partly cancels the two gradient noise terms near equal weights. Changing q from 1 to 2 reduces that cancellation and raises $\sigma^2(q)$. The curvature $\mathcal{H}(q)$ also changes, and the exact risk formula accounts for both changes.

C.5 What independent minibatches can measure

Eq.(C.6) also indicates how to remove the same-sample covariance term from the expectation of a measurement. At a fixed model and timestep, draw two independent minibatches. Use superscripts (1) and (2) to indicate which minibatch supplies each gradient, and form both directions with the same weights chosen before either minibatch is drawn. To use the two batches symmetrically, define

$$\widetilde{a}_H^w = \frac{1}{2}\left[\langle g_H^{(1)}, d^{w,(2)}\rangle + \langle g_H^{(2)}, d^{w,(1)}\rangle\right]. \tag{C.28}$$

The tilde identifies this cross-minibatch quantity, as distinct from a_H^w, which uses a single minibatch. Independence gives

$$\mathbb{E}[\widetilde{a}_H^w] = \langle G_H, w_L G_L + w_H G_H\rangle. \tag{C.29}$$

Its value on a single pair of minibatches still varies with sampling. Computing it requires the two sets of gradient vectors, or their cross-products; separate summaries of gradient norms and cosines do not contain that information.

Prior work has examined why gradient conflict need not predict multi-task generalization (Mueller et al., 2024), and why denoising weights can affect gradient variance and statistical efficiency differently (Zhang et al., 2026). Here the covariance identity and denoising construction connect the minibatch directional criterion directly to population gradients and stationary excess risk.

The main paper uses minibatch gradients and test updates to examine local frequency interaction, and FID comparisons to assess generation after training. The construction explains why both kinds of evidence are needed. The independent-minibatch estimator is a possible extension of the diagnostic; it is not used in the reported experiments.

D Diagnostic Protocol and Supporting Statistics

The diagnostics measure band gradients in the shared input block, error changes after a low-frequency test step, and gradient contributions after baseline and FAL training. Each measurement uses the definitions above with the parameter block, timestep, and aggregation specified below.

D.1 Condition test in the shared input block

At each saved checkpoint, we freeze the denoiser and select its shared input weight block `init_conv.weight`. For each checkpoint and timestep, we calculate g_L and g_H on several fixed diagnostic minibatches. We first calculate R, ρ, and ϕ_H separately within each minibatch, then take the median of each statistic across minibatches. The evaluated timesteps are

$\{100, 250, 400, 550, 700, 850\}$. Both the L2 and L1 band-loss tests use the same spectral cutoff $c = 0.35$. The summary at $t = 250$ uses the 14 post-warmup checkpoints shown for each loss in the main paper, excluding the 10K and 20K checkpoints. Its two median columns take the median of the 14 checkpoint statistics, and its last column counts checkpoints with negative ϕ_H.

Table D.1: Condition test at $t = 250$ over 14 checkpoints after warmup. Each checkpoint statistic aggregates multiple diagnostic minibatches.

Objective	Ckpts.	Median R	Median ϕ_H	$\phi_H < 0$
L2	14	28.91	-0.0046	14/14
L1	14	29.25	-0.0059	14/14

Since ϕ_H and M_H share a sign for a nonzero joint direction, the last column counts checkpoints satisfying the high-frequency increase condition in this input block.

D.2 Error change across output frequencies

The gradient test concerns joint updates. To measure the effect of the low-frequency direction alone, we record error changes across output frequencies after a test step. We use checkpoints from early (10K), middle (200K), and late (300K) training and 41 diffusion timesteps from 500 through 900 in increments of 10. At each measurement, we take one fixed small test step along $-\widehat{g}_L$, the negative normalized low-frequency gradient. Let f index a radial output-frequency bin and let $\mathcal{E}_f$ be the squared residual error in that bin. The relative change in that bin is

$$\Delta_{L\to f} = \frac{\mathcal{E}_f(\vartheta - \eta\widehat{g}_L) - \mathcal{E}_f(\vartheta)}{\mathcal{E}_f(\vartheta)}. \tag{D.1}$$

The saved diagnostic divides radial frequency from 0 to 0.6 cycles/pixel into 24 bins of width 0.025 cycles/pixel. The common cutoff $c = 0.35$ corresponds to 0.175 cycles/pixel here, separating seven low-frequency bins from 17 high-frequency bins. For each stage, the stored array contains four measurements at each of the 41 timesteps. The destructive-transfer rate pools the signs of all stored measurements in the relevant bins and reports the fraction with $\Delta_{L\to f} > 0$. The test-step size is held fixed across the three stages.

Table D.2: Fraction of frequency-bin measurements whose error rises after the low-frequency test step. Here pp means percentage points.

Stage	Low bins	High bins	High–low gap
Early (10K)	3.1%	46.5%	43.4 pp
Middle (200K)	14.0%	55.3%	41.3 pp
Late (300K)	11.8%	51.0%	39.2 pp

Figure D.1 resolves the measurements in Table D.2 by noise timestep and frequency bin for each training stage.

D.3 Comparison after baseline and FAL training

We compare a baseline model and a separately trained FAL model, each using its own band gradients at its learned parameters. Their normalized high-frequency contributions, denoted ϕ_H^{Base} and ϕ_H^{FAL}, follow Eqs.(B.13) and (B.18), respectively. At each timestep, define $\Delta\phi_H(t) = \phi_H^{\mathrm{FAL}}(t) - \phi_H^{\mathrm{Base}}(t)$. In Table D.3, the expected direction is $\Delta\phi_H > 0$ at low noise and $\Delta\phi_H < 0$ at high noise; the negative-margin counts are ordered baseline $\to$ FAL.

All 27 timesteps with a negative baseline margin move upward in the low-noise region, and 22 cross zero. These changes reflect both the FAL weights and the gradients learned during training.

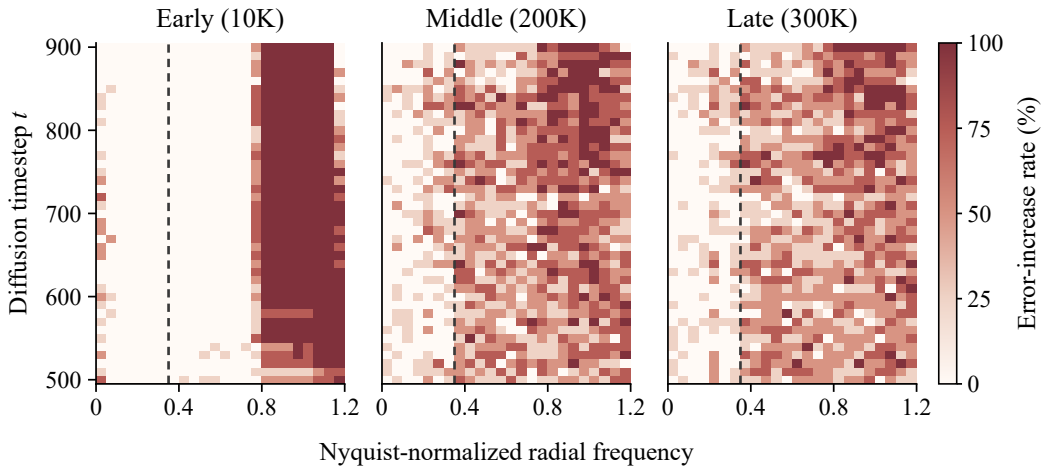

Figure D.1: Frequency-resolved error increases after the low-frequency test update. Each cell reports the percentage of four measurements with $\Delta_{L\to f} > 0$ at one diffusion timestep and frequency bin. The dashed line marks $c = 0.35$. The three panels use the same measurements summarized in Table D.2.

Table D.3: Change in high-frequency gradient contribution between the paired models. Expected direction counts upward changes at low noise and downward changes at high noise. The last column counts timesteps with a negative high-frequency margin.

Timestep region	Mean $\Delta\phi_H$	Expected direction	Negative margins
1–499	$+5.43\times10^{-4}$	73.9%	$27 \to 5$
500–999	-1.19×10^{-2}	98.8%	–

E SpecDiffusion Implementation

SpecDiffusion applies FAL to the denoising residual. Configurations with FVA also provide local frequency views at the input; the component study evaluates the two changes separately.

E.1 Fourier projections and FAL

Let $z \in \mathbb{R}^{C\times H\times W}$ be a spatial tensor with C channels, height H, and width W. We apply the two-dimensional Fourier transform independently to each channel. Denote this transform by $\mathcal{F}$ and its inverse by $\mathcal{F}^{-1}$. Let ω be a spatial frequency coordinate and let $\omega_N = 0.5$ cycles/pixel be the Nyquist frequency along one axis. The normalized radial frequency is $\nu(\omega) = \|\omega\|_2/\omega_N$. The fixed masks are

$$\chi_L(\omega) = \mathbf{1}[\nu(\omega) \le c], \qquad \chi_H(\omega) = 1 - \chi_L(\omega), \tag{E.1}$$

where $\mathbf{1}[\cdot]$ equals one when its condition holds and zero otherwise. For $B \in \{L, H\}$, the projection is

$$P_B z = \mathcal{F}^{-1}(\chi_B \odot \mathcal{F}(z)), \tag{E.2}$$

where $\odot$ multiplies the mask and Fourier coefficients element by element. The cutoff is 0.175 cycles/pixel before Nyquist normalization. We use the recommended cutoff $c = 0.35$ throughout all resolutions, datasets, backbones, and prediction targets.

Apply these projections to the residual r_θ from Section A. FAL weights the projected squared errors:

$$\mathcal{L}_{\text{FAL}} = \frac{1}{2}\mathbb{E}\left[w_L(t)\|P_L r_\theta\|_2^2 + w_H(t)\|P_H r_\theta\|_2^2\right]. \tag{E.3}$$

Here the expectation is over data, noise, and diffusion timesteps. With T total diffusion timesteps and $t \in \{0, \dots, T-1\}$, use $\bar{t} = t/(T-1)$ as before:

$$w_L(t) = 1 + \bar{t}, \qquad w_H(t) = 2 - \bar{t}. \tag{E.4}$$

Each projected residual has the original tensor shape. We use the same mean reduction over batch, channel, and spatial positions as the matched baseline. This introduces a common scale for both bands; the formulas retain the squared-norm convention. For the L1 component comparison, we retain the masks and weights and replace each band's squared error by its absolute error. Unlike the squared loss, these two losses need not sum to the unprojected L1 loss.

E.2 Frequency View Adapter

The Frequency View Adapter (FVA) supplies local low- and high-frequency views of the noisy input x_t. These views use spatial filtering, whereas FAL uses the global Fourier projection above. Let $\mathcal{B}_{\mathrm{blur}}$ be a fixed channel-wise 5×5 binomial blur with reflection padding. Define

$$Q_L x_t = \mathcal{B}_{\mathrm{blur}}(x_t), \qquad Q_H x_t = x_t - \mathcal{B}_{\mathrm{blur}}(x_t). \tag{E.5}$$

The blur's spatial kernel K_{blur} is

$$K_{\mathrm{blur}} = \frac{1}{256} \begin{bmatrix} 1 \\ 4 \\ 6 \\ 4 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 4 & 6 & 4 & 1 \end{bmatrix}. \tag{E.6}$$

For a convolutional denoiser, let E_0, E_L, E_H be the input embedding maps for the original, low-, and high-view inputs. Their representations are

$$h_0 = E_0(x_t), \quad h_L = E_L(Q_L x_t), \quad h_H = E_H(Q_H x_t). \tag{E.7}$$

A residual branch combines these representations. Let γ be its scale, let square brackets concatenate tensors along channels, and let $\mathrm{Conv}_{1\times1}$ be its fusion convolution. The adapted representation is

$$\widetilde{h}_0 = h_0 + \gamma \ \mathrm{Conv}_{1\times1}([h_0, h_L, h_H]). \tag{E.8}$$

The branch is initialized so that its added output is zero at the start of training. For token denoisers, the three spatial views are embedded at patch level and fused before the later Transformer blocks. For latent models, both view operators act on the noisy spatial latent tensor.

E.3 Training sequence

Each training iteration follows these steps:

1. sample a diffusion timestep and form the noisy image or latent tensor;
2. when FVA is used, construct its two local input views;
3. obtain the denoiser prediction and residual;
4. apply the two fixed Fourier projections to that residual;
5. evaluate Eq.(E.3) and update the model with its original optimizer.

The prediction target, prediction head, and sampling procedure remain those of the matched baseline.

F Data, Training, and Evaluation Protocols

Each baseline/SpecDiffusion pair shares its data pipeline, training budget, and evaluation protocol. The interventions are FAL and, where indicated, FVA; the settings below specify the controls and the differences between model families.

F.1 Datasets and preprocessing

We use the official releases of CIFAR-100, CelebA, FFHQ, LSUN-Church, and ImageNet (ILSVRC2012), with the standard training split of each dataset. The shared data pipeline includes resolution, decoding, normalization, cropping, and augmentation from each model family's project configuration.

The dataset names in the result tables specify image resolution; CIFAR-100 uses 32×32 images. Lightning-DiT uses its official VA-VAE encoder and velocity-prediction target.

F.2 Matched training configurations

ADM, LDM, and Lightning-DiT follow their released project configurations. Their paired runs retain the denoiser backbone, optimizer, learning-rate schedule, minibatch construction, diffusion parameterization, conditioning, and exponential moving average of parameters (EMA). Lightning-DiT uses FAL alone with its official input pipeline.

UNet-DDPM and Pixel-DiT start from the standard DDPM implementation, with the same denoiser, optimizer, learning-rate schedule, batch size, and checkpoint selection within each pair. Table F.1 lists the resources, components, and number of independent runs.

Table F.1: Training resources and repetitions. All GPUs are A100-80GB.

Family	GPUs/run	Realization	Independent runs
UNet-DDPM	1	FAL+FVA	3
Pixel-DiT	1	FAL+FVA	3
ADM	1	FAL+FVA	3
LDM	2	FAL+FVA	3
Lightning-DiT	8	FAL	1

In the main comparisons, Fréchet Inception Distance (FID) is reported as mean and standard deviation over the independent runs listed above, except for Lightning-DiT's single-run result.

F.3 Sampling for the main comparisons

Table F.2 records each family's reverse sampler, number of steps, and sample count for the main FID comparisons. The FID implementation is also shared within each pair.

Table F.2: Sampling used for the main matched comparisons.

Family	Sampler	Steps	Samples
UNet-DDPM	DDPM	1000	50K
Pixel-DiT	DDIM	200	50K
ADM	DDIM	250	50K
LDM	DDIM	250	50K
Lightning-DiT	DDIM	250	50K

G Additional Quantitative Results

The tables collect the main paper's matched and component results for reference, followed by the full checkpoint results of the weight-schedule study.

G.1 Complete matched FID comparisons

Tables G.1 and G.2 list the paired FID results for image-pixel and spatial-latent denoisers, respectively, using the protocols above.

Table G.1: Pixel-space matched FID results; lower is better.

Family	Dataset	Baseline	SpecDiffusion
UNet-DDPM	FFHQ128	13.24 ± 0.18	$\mathbf{11.07 \pm 0.16}$
UNet-DDPM	CelebA128	11.38 ± 0.14	$\mathbf{9.88 \pm 0.12}$
UNet-DDPM	LSUN-Church128	14.77 ± 0.20	$\mathbf{13.43 \pm 0.18}$
ADM	CIFAR-100	6.295 ± 0.07	$\mathbf{5.828 \pm 0.06}$
ADM	CelebA64	10.253 ± 0.11	$\mathbf{8.042 \pm 0.09}$
ADM	FFHQ64	7.830 ± 0.09	$\mathbf{6.801 \pm 0.08}$
Pixel-DiT	CIFAR-100	16.94 ± 0.21	$\mathbf{13.78 \pm 0.19}$
Pixel-DiT	CelebA64	14.31 ± 0.19	$\mathbf{13.30 \pm 0.17}$

Table G.2: Spatial-latent matched FID results; lower is better.

Family	Dataset	Baseline	SpecDiffusion
LDM	CelebA128	19.54 ± 0.22	$\mathbf{19.10 \pm 0.20}$
LDM	LSUN-Church256	10.45 ± 0.12	$\mathbf{9.14 \pm 0.11}$
LDM	FFHQ256	15.23 ± 0.19	$\mathbf{14.39 \pm 0.03}$
Lightning-DiT	ImageNet256	2.163	**2.028**

The latent-space results evaluate spatial reweighting on latent tensors; the direct frequency-gradient measurements use the pixel-space diagnostic described above.

G.2 Loss and input-adapter components

The FFHQ 256^2 study trains each configuration for up to 400K steps. Table G.3 lists its best reported FID and the training step associated with that value. L1 and L2 name the within-band error norm; FAL changes the loss, and FAL+FVA also adds the input adapter.

Table G.3: FFHQ 256^2 component study over 400K training steps. FID is mean $\pm$ standard deviation over three independent runs.

Loss	Configuration	Best FID	Step
L1	Baseline	14.04 ± 0.12	260K
L1	FAL	12.47 ± 0.45	140K
L1	FAL+FVA	$\mathbf{12.09 \pm 0.33}$	90K
L2	Baseline	18.46 ± 0.54	250K
L2	FAL	$\mathbf{15.75 \pm 0.26}$	230K
L2	FAL+FVA	15.89 ± 0.81	170K

FAL alone improves the mean FID under both error norms; the additional effect of FVA differs between L1 and L2.

G.3 Weight schedules

The CIFAR-100 schedule study compares linear, cosine, quadratic, and step schedules for band priority. Use the same normalized timestep $\bar{t} = t/(T-1)$ as in the main paper; $\bar{t} = 1$ is the high-noise end. The four weight pairs are

$$(w_L, w_H)_{\text{linear}} = (1 + \bar{t}, 2 - \bar{t}), \tag{G.1}$$

$$(w_L, w_H)_{\text{cosine}} = \left(1 + \tfrac{1}{2}[1 - \cos(\pi\bar{t})], 1 + \tfrac{1}{2}[1 + \cos(\pi\bar{t})]\right), \tag{G.2}$$

$$(w_L, w_H)_{\text{quadratic}} = \left(1 + \bar{t}^2, 1 + (1 - \bar{t})^2\right), \tag{G.3}$$

$$(w_L, w_H)_{\text{step}} = (1 + \mathbf{1}[\bar{t} > 1/2], 1 + \mathbf{1}[\bar{t} \le 1/2]). \tag{G.4}$$

The linear variant is Eq.(E.4). At each checkpoint, all rows use 50-step DDIM sampling and 10K generated images for FID. Schedule choices are therefore compared within Table G.4, whose evaluator differs from the main comparisons in Table F.2.

Table G.4: CIFAR-100 schedule study. FID uses 10K samples and 50-step DDIM.

Method	10K	20K	30K	40K
Baseline	328.78	92.66	17.26	13.37
FAL-linear	338.00	103.27	17.72	12.01
FAL-cosine	332.97	102.30	19.28	12.22
FAL-quadratic	327.48	102.86	17.65	12.17
FAL-step	332.27	94.22	27.17	19.14
Full (linear FAL+FVA)	335.83	102.36	**16.55**	**11.89**

The columns mark training checkpoints. At the fixed 40K endpoint, the three smooth FAL schedules have relatively small FID differences and improve on the baseline, while the tested step schedule does not. The smooth schedules have higher FID than the baseline at 20K, so the endpoint gains arise after further training. We use the linear rule as a simple schedule for varying band priority.