ON HIGH-FREQUENCY COLLAPSE IN SPECTRAL VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern vision architectures often emphasize low-frequency representations as depth increases (Yun et al., 2023; Rahaman et al., 2019). In spectral token mixers, this tendency can suppress high-frequency (HF) information—edges, textures, and fine local structure—that is discriminative for detail-dependent recognition. We study this architecture-dependent failure mode, which we call *high-frequency collapse*, and introduce **HF-Skip**: a frequency-selective pathway that applies the learned spectral gate to low-frequency coefficients while routing the upper-frequency band around it. Controlled band-filtered datasets isolate frequency content from dataset size, labels, and image identity. On ImageNet-100 at split radius $r = 0.12$, HF-Skip improves the HF-only arm by $+12.21 \pm 0.25$ points but the matched LF-only arm by only $+0.67 \pm 0.70$. Parameter-matched controls on DTD further show that HF-Skip gains $+8.90$ points, compared with $+6.78$ for a full-spectrum bypass, $+4.15$ for a random-bin bypass, $+0.53$ for an inverted frequency mask, and $+0.00$ for a generic spatial residual. On standard benchmarks, HF-Skip improves SpectFormer by $+1.27$ points on ImageNet-100 and $+8.83$ on DTD, generalizes to GFNet, and adds only 0.009% FLOPs. Full ImageNet-1K accuracy is tied (79.674 vs. 79.678), delimiting the benefit to regimes where fine detail matters rather than claiming a universal gain. Together, these results identify HF collapse as a task-dependent bottleneck and show that frequency-selective routing preserves useful fine detail with negligible computational overhead.

1 INTRODUCTION

Vision models rely fundamentally on high-frequency (HF) information, such as edges, textures, and fine-grained spatial details, to distinguish objects, recognize patterns, and make accurate predictions (Azad et al., 2023; Ma et al., 2025). Yet, empirical observation reveals that modern Vision Transformers (ViTs), particularly those incorporating spectral processing, progressively bias toward low-frequency (LF) representations as network depth increases (Bai et al., 2022). While this bias is typically benign on object-centric benchmarks where global shape dominates, it becomes critically limiting when tasks depend on fine-grained texture or local structure (Azad et al., 2023).

The tendency of deep neural networks to favor low-frequency components is well-documented. In Convolutional Neural Networks (CNNs), this manifests through pooling operations, strided convolutions, and the hierarchical aggregation of local features into global representations (Rahaman et al., 2019; John Xu et al., 2020). This *spectral bias*, where networks learn low-frequency functions more readily than high-frequency ones, has been linked to implicit regularization and improved generalization.

In Vision Transformers (ViTs) (Dosovitskiy et al., 2021), a similar low-frequency emphasis emerges through different mechanisms. The patch embedding operation imposes spatial discretization, and the global self-attention mechanism, while theoretically capable of capturing arbitrary spatial frequencies, acts as a low-pass filter by preferentially connecting tokens with similar content to smooth representations (Park & Kim, 2022; Rahaman et al., 2019). This implicit smoothing intensifies with depth as information is repeatedly aggregated across the global token sequence.

We distinguish between *low-frequency bias* and the more severe behavior we call *high-frequency collapse*. Low-frequency bias is a broad preference for smoother functions and can be beneficial. We use HF collapse as an operational, task-dependent label for cases in which recognition suffers

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.

because a learned spectral pathway does not preserve useful fine-scale information. It is not intended to describe a universal property of spectral models or every reduction in HF energy. Our central claim is empirical and intervention-based: learned spectral gating can become a bottleneck for fine-scale information, and selectively routing the affected band around that bottleneck improves performance when those frequencies are discriminative.

HF collapse should be most consequential when task-relevant information depends on fine-scale structure. On the Describable Textures Dataset (DTD), SpectFormer (Patro et al., 2023) reaches 21.44% accuracy (3-seed mean), below an attention-only ViT at 25.48%, whereas HF-Skip reaches 30.27% ($+8.83$ points). More importantly, the controlled ImageNet-100 intervention changes only the surviving frequency band: HF-Skip gains 12.21 ± 0.25 points on HF-only inputs at $r = 0.12$, compared with 0.67 ± 0.70 on the matched LF-only inputs. This within-dataset contrast, rather than an assumed binary distinction between "texture" and "object" datasets, motivates the frequency-specific claim.

Our contributions are five-fold. (1) We identify and measure *high-frequency collapse* as an architecture-dependent bottleneck in spectral vision models, distinct from the general low-frequency bias of gradient-based learning. (2) We evaluate HF-Skip across representative spectral architectures and isolate the frequency dependence of its empirical gain using matched band-filtered datasets, where frequency content is manipulated while image identity, labels, splits, and sample count remain fixed. (3) We propose **HF-Skip**, a minimal frequency-selective routing modification that preserves selected coefficients outside the learned spectral gate. (4) We test the frequency-specific explanation against parameter-matched alternatives, including full-spectrum, random-bin, inverted-mask, and generic spatial residual controls. (5) We characterize the method's scope rather than assuming universal improvement: it yields large gains on texture- and HF-dependent inputs, ties the baseline on full ImageNet-1K, has a dataset-dependent threshold, and exposes a tunable sensitivity to HF corruption.

2 Related Work

Deep neural networks exhibit a well-documented *spectral bias*, i.e., a tendency to learn low-frequency components of target functions before high-frequency ones. This behavior has been linked to the eigenstructure of the Neural Tangent Kernel and the smoothness induced by gradient-based optimization, and has been observed across MLPs, CNNs, and transformers (Rahaman et al., 2019; Xu et al., 2020). While this low-frequency preference often acts as an implicit regularizer favoring smooth solutions, it can be detrimental when task-relevant information lies in fine spatial details such as textures or edges (Basri et al., 2019; Cao et al., 2020). Architecturally, both CNNs and Vision Transformers reinforce this bias: CNNs through pooling, strided convolutions, and hierarchical aggregation, and transformers through patch embeddings and attention-based smoothing, leading to frequency attenuation and representation collapse in deeper layers (Geirhos et al., 2022; Raghu et al., 2022; Dong et al., 2023). Prior work largely attributes this behavior to optimization or attention dynamics; our untrained-model analysis instead asks how an explicit frequency-selective route changes activation statistics before task optimization. Recent work has explored *spectral vision architectures* that replace or augment attention with frequency-domain token mixing. GFNet demonstrates that learnable Fourier-domain filters can serve as effective attention substitutes (Rao et al., 2021), while AFNO introduces adaptive sparsification and channel mixing in the frequency domain (Guibas et al., 2022). FNet further shows that parameter-free Fourier transforms can suffice for token interaction (Lee-Thorp et al., 2022), and hybrid models such as SpectFormer combine spectral and attention-based processing across layers (Patro et al., 2023). Despite strong empirical performance, the internal frequency behavior of these models remains underexplored. We study when a learned spectral gate becomes an information bottleneck and evaluate a direct intervention: route a selected frequency band around the gate while leaving the rest of the spectral mixer unchanged.

Preserving high-frequency information is critical for fine-grained recognition tasks such as texture classification, medical imaging, and remote sensing (Cimpoi et al., 2015; Shen et al., 2017). Existing approaches attempt to mitigate high-frequency loss through spatial-domain techniques including wavelet embeddings, Laplacian filtering, and localized attention (Liu et al., 2021; Yao et al., 2022). Separately, the problem of *oversmoothing*, where repeated aggregation drives representations toward low-dimensional subspaces, has been studied in graph neural networks and transformers (Oono & Suzuki, 2021; Dong et al., 2023). HF-Skip differs in where it intervenes: it changes information

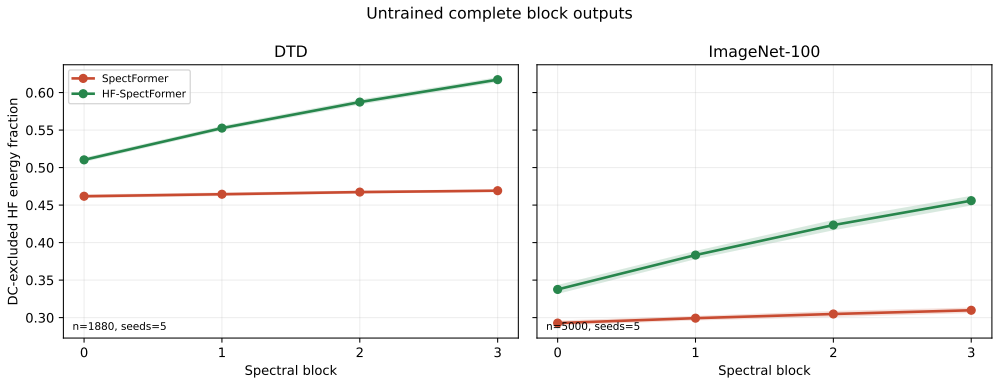

Figure 1: **Frequency distribution of untrained spectral-block outputs.** We report the DC-excluded high-frequency energy fraction at the complete output of spectral blocks 0–3 for matched SpectFormer and HF-SpectFormer initializations. Curves show mean and standard deviation over five initialization seeds, with energies aggregated over all 1,880 DTD validation images and 5,000 ImageNet-100 validation images. HF-Skip changes the frequency distribution of untrained block outputs.

routing inside a spectral mixer, and our controlled experiments test whether the benefit depends specifically on which frequency band is protected.

Prior work has documented high-frequency attenuation in vision transformers (Bai et al., 2022). Our contribution is to study this behavior in learned spectral gates and to test frequency-selective routing through matched band-filtered datasets and parameter-matched architectural controls.

3 Characterizing a High-Frequency Bottleneck

We define the measured quantities and reporting conventions used throughout the paper. For a real-valued feature map $X \in \mathbb{R}^{H\times W}$, a spectral block computes the one-sided Fourier spectrum $\hat{X} = \mathrm{rfft2}(X)$, multiplies it by a learned complex gate K, and returns to the spatial domain with $\mathrm{irfft2}$. We construct radial coordinates directly on the stored, unshifted $\mathrm{rfft2}$ layout and account for the multiplicity of interior frequencies. Unless otherwise stated, the reporting radius $r = 0.5$ divides the grid into LF and HF regions. After excluding the DC coefficient, we report the HF energy fraction $\rho_{\mathrm{HF}} = E_{\mathrm{HF}}/(E_{\mathrm{LF}} + E_{\mathrm{HF}})$. This reporting convention is distinct from the routing threshold τ, which determines which coefficients bypass the learned gate. We use *HF collapse* as an operational, task-dependent label for cases in which a learned spectral pathway becomes a bottleneck for predictive fine-scale information; it neither treats every reduction in HF energy as harmful nor asserts a specific inverse-FFT reconstruction mechanism. At matched random initialization, HF-Skip produces a higher DC-excluded HF energy fraction than SpectFormer at the complete outputs of spectral blocks 0–3 on DTD and ImageNet-100 (Figure 1), with the ordering holding across all five initialization seeds. This result establishes only that routing changes the distribution of untrained activations; evidence for frequency-specific performance attribution comes from the controlled band-filtered datasets and parameter-matched interventions in Sections 5.1 and 6.1.

4 Methodology

4.1 Frequency Distribution at Initialization

Before introducing the proposed architecture, we measure how its routing choice changes activations prior to task optimization. We compare parameter-matched SpectFormer and HF-SpectFormer under five matched random initializations, using all 1,880 DTD validation images and all 5,000 ImageNet-100 validation images. For each complete block output, we report the DC-excluded HF energy fraction using the correct unshifted one-sided frequency layout and Parseval multiplicity.

HF-Skip has a higher HF fraction in all 40 dataset–seed–block comparisons. From block 0 to block 3, the mean difference between HF-SpectFormer and SpectFormer grows from 0.0485 to 0.1480 on

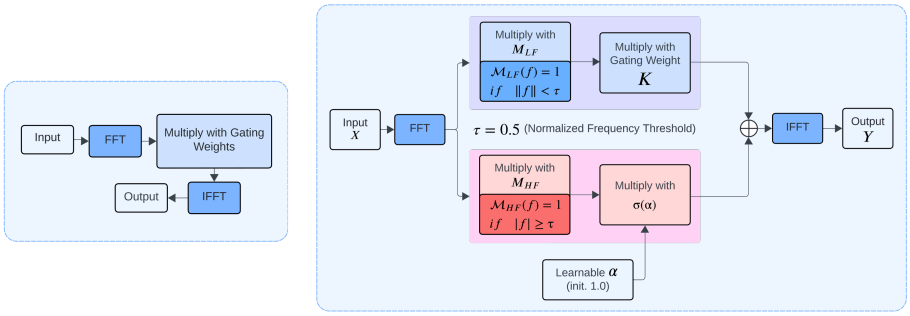

Figure 2: Left: Spectral block mechanism of the original SpectFormer. Right: Spectral block mechanism of HF-Skip.

DTD and from 0.0448 to 0.1461 on ImageNet-100. These measurements describe a consistent change in the distribution of untrained block outputs; they do not establish that the additional HF energy is task-relevant or explain the trained accuracy gain. We test those questions with the controlled interventions in Section 6.

4.2 HF-Skip: Frequency-Selective Residuals

To prevent selected coefficients from depending entirely on the learned spectral gate, we introduce **HF-Skip**, an architectural block that provides an identity route for the upper-frequency region. We integrate HF-Skip into SpectFormer (**SpectFormer + HF-Skip**) and also test it in GFNet. The standard Spectral Gating layer operation can be written as:

$$Y = \text{IFFT}\big(K \odot \text{FFT}(X)\big) \tag{1}$$

where K represents the learnable gating weights.

The **HF-Skip** modification applies gating *selectively* to low-frequency components while preserving high-frequency components:

$$Y = \text{IFFT}\big(K \odot \mathcal{M}_{LF} \odot \hat{X} + \sigma(\alpha) \cdot \mathcal{M}_{HF} \odot \hat{X}\big) \tag{2}$$

where $\hat{X} = \text{FFT}(X)$ and:

- K is the learnable complex gating weight.
- $\mathcal{M}_{LF}(f) = \mathbf{1}[\|f\| < \tau]$ and $\mathcal{M}_{HF} = 1 - \mathcal{M}_{LF}$ are complementary frequency masks, precomputed once per spatial resolution and shared across channels and batches.
- α is a learnable scalar (initialized to 1.0), passed through sigmoid $\sigma(\cdot)$ to bound the skip ratio.
- τ defines the normalized routing threshold. We use $\tau = 0.5$ as a fixed cross-dataset default in the headline comparisons and report full, dataset-specific sweeps in Section 6.3.

Crucially, the gating weight K is applied **only to the LF region**; coefficients in the complementary region retain an identity pathway rather than being determined solely by K.

HF-Skip operates by explicitly **decoupling low- and high-frequency signal pathways** within the spectral mixer. Low-frequency components ($< \tau$) continue to pass exclusively through the multiplicative gate K. This preserves the intended function of the spectral token mixer: modeling global structure and long-range dependencies, which are predominantly encoded in low-frequency modes. High-frequency components ($> \tau$), by contrast, are routed through an additive identity pathway. By reinjecting the original high-frequency spectrum prior to the inverse FFT, HF-Skip ensures that fine-grained details(such as edges and textures) bypass the multiplicative attenuation imposed by the randomly initialized gate. This design changes the layer's default behavior. Rather than making upper-band coefficients depend entirely on a small learned gate, the identity route preserves them by default. Per-frequency modulation in the bypassed region is reduced to a block-level scale α, while later layers retain capacity to combine and reshape the resulting features. The intervention therefore makes strong HF attenuation an explicit routing choice instead of an unavoidable consequence of the gated path.

Table 1: Performance comparison across texture and natural image benchmarks (3-seed mean ± std for SpectFormer-S and HF-SpectFormer; single-run for ViT-Small, GFNet-Small, and Imagenette). HF-SpectFormer consistently improves accuracy by preserving high-frequency information critical for texture and fine-grained recognition. Per-seed numbers are in Appendix A.

	DTD			ImageNet-100			Imagenette	
Model	Top-1	Top-5	Δ	Top-1	Top-5	Δ	Top-1	Δ
ViT-Small	25.48	53.72	+4.04	79.70	94.40	−2.79	73.83	−2.22
GFNet-Small	22.50	55.00	+1.06	81.56	95.22	−0.93	74.30	−1.75
SpectFormer-S	21.44 ± 1.51	49.08 ± 1.94	baseline	82.49 ± 0.03	95.55 ± 0.15	baseline	76.05	baseline
HF-SpectFormer	**30.27 ± 0.08**	**61.52 ± 0.56**	**+8.83**	**83.76 ± 0.44**	**95.97 ± 0.27**	**+1.27**	**79.49**	**+3.44**

5 EXPERIMENTS

We evaluate our hypothesis and proposed remedy on three image classification benchmarks spanning different frequency characteristics. **ImageNet-100** (Tian et al., 2020), a 100-class subset of ImageNet-1k with ∼130k training and 5k validation images, serves as our primary benchmark for general object recognition. **Imagenette** (Howard, 2019), a lightweight 10-class subset of ImageNet with ∼9k training images, is used to study convergence behavior. Finally, we evaluate on the **Describable Textures Dataset (DTD)** (Cimpoi et al., 2014), containing 5,640 images across 47 texture categories. Since DTD relies heavily on high-frequency texture cues, it is particularly suitable for testing whether frequency-selective routing benefits detail-dependent recognition.

To evaluate the proposed modification across representative architectures, we compare four models with comparable parameter counts at the DeiT-Small scale. **ViT-Small** (Dosovitskiy et al., 2021) serves as a pure-attention baseline with an implicit low-frequency bias but no explicit spectral bottleneck. **GFNet-Small** (Rao et al., 2021) provides a pure spectral baseline using learnable FFT-domain filters without spectral gating. **SpectFormer-S** is the original hybrid architecture combining spectral token mixing and self-attention, and serves as the primary subject of our HF collapse analysis. Finally, **HF-SpectFormer (Ours)** extends SpectFormer with frequency-selective skip connections to mitigate high-frequency collapse. All models are comparable in scale: GFNet-Small (15.64M, 5.80 GFLOPs), SpectFormer-S (19.68M, 7.59 GFLOPs), HF-SpectFormer (19.68M, 7.59 GFLOPs), and ViT-Small (21.70M, 8.49 GFLOPs). HF-Skip adds only +0.009% FLOPs (+0.69 MFLOPs). Details are in Table 18 in the Appendix. All models are trained from scratch using a unified training recipe implemented in `timm`. We use the AdamW optimizer (weight decay 0.05) with a cosine annealing learning-rate schedule (initial LR 5×10^{-4}, 5-epoch warmup). Data augmentations include RandomResizedCrop, Mixup ($\alpha = 0.8$), CutMix ($\alpha = 1.0$), and RandomErasing. Training is conducted for 150 epochs on ImageNet-100 and 100 epochs on Imagenette and DTD.

Standard Benchmarks HF-SpectFormer improves performance on standard object recognition tasks without introducing regressions. As shown in Table 1, HF-SpectFormer achieves the highest accuracy across all evaluated architectures. On ImageNet-100, it improves over the baseline SpectFormer by **+1.27%** (3-seed mean). Although object recognition is largely shape-driven (low-frequency), preserving edge and boundary information still helps distinguish visually similar classes. On Imagenette, the performance gap widens to **+3.44%** (79.49% vs. 76.05%) suggesting that in limited-data regimes, where models cannot rely on large sample sizes to learn robust shape priors, explicit preservation of high-frequency features provides a stronger discriminative signal.

Texture Analysis To test our hypothesis that HF collapse harms detail-dependent tasks, we evaluate on the Describable Textures Dataset (DTD). Table 1 shows that SpectFormer (21.44%) falls below both the purely spectral GFNet (Rao et al., 2021) (22.50%) and the attention-only ViT (25.48%). **HF-SpectFormer** reaches **30.27%** accuracy (3-seed mean), improving by **+8.83** points over the baseline. The larger improvement on DTD than on ImageNet-100 is consistent with texture recognition depending more heavily on fine spatial structure, but cross-dataset differences alone cannot establish frequency-specific causality. We test that attribution directly next.

Table 2: Matched band-filtered datasets. "Energy" is the fraction of the original image's spectral energy retained by each frequency mask before per-channel mean and variance restoration. Δ is HF-Skip minus SpectFormer top-1 accuracy. The final column reports the ratio of the point-estimate gains on the HF-only and LF-only arms at the same split.

Dataset	Split r	Energy	Arm	SpectFormer	HF-Skip	Δ	HF/LF gain
DTD	0.10	19.00%	HF-only	16.13 ± 1.16	25.28 ± 0.55	$+9.15$	$3.15\times$
		81.00%	LF-only	19.01 ± 0.78	21.91 ± 0.47	$+2.91$	
	0.15	13.95%	HF-only	11.79 ± 1.26	23.26 ± 0.65	$+11.47$	$2.81\times$
		86.05%	LF-only	20.94 ± 1.21	25.02 ± 0.76	$+4.08$	
	0.20	10.54%	HF-only	11.44 ± 0.84	20.59 ± 0.53	$+9.15$	$1.86\times$
		89.46%	LF-only	22.13 ± 0.54	27.04 ± 0.25	$+4.91$	
	0.30	6.19%	HF-only	10.16 ± 0.37	18.56 ± 0.80	$+8.40$	$1.32\times$
		93.81%	LF-only	22.07 ± 0.75	28.44 ± 1.00	$+6.37$	
	0.50	1.94%	HF-only	9.52 ± 0.23	12.80 ± 0.35	$+3.28$	$0.48\times$
		98.06%	LF-only	22.15 ± 0.84	28.95 ± 0.66	$+6.81$	
ImageNet-100	0.12	12.26%	HF-only	61.00 ± 0.30	73.21 ± 0.16	$+12.21 \pm 0.25$	$18.3\times$
		87.74%	LF-only	72.68 ± 0.27	73.35 ± 0.53	$+0.67 \pm 0.70$	
	0.50	1.71%	HF-only	40.22 ± 0.59	47.81 ± 1.22	$+7.59 \pm 1.66$	$3.4\times$
		98.29%	LF-only	78.01 ± 0.53	80.21 ± 0.44	$+2.21 \pm 0.53$	

5.1 Controlled Band-Filtered Datasets

We construct matched LF-only and HF-only versions of DTD and ImageNet-100. For every image, we compute a 2D DFT, apply a complementary radial mask at split radius r, invert the result, and restore the original per-channel mean and variance. The two versions preserve image identity, label, split, resolution, sample count, and training recipe; only the surviving frequency band changes. Table 2 reports three-seed results.

The cleanest contrast is ImageNet-100 at $r = 0.12$: the method gains 12.21 ± 0.25 points when only the HF arm remains, but 0.67 ± 0.70 on the matched LF-only arm. Thus the improvement on an object-recognition dataset is not contradictory: object images are LF-dominated overall, yet their smaller HF component can still carry discriminative boundaries and texture. DTD shows a broader transition rather than a binary split. HF-only gains dominate for $r = 0.10$–0.30, while the ordering reverses at $r = 0.50$; this indicates that DTD's task-relevant structure extends into the mid-frequency range and that conclusions depend on the chosen radial partition.

Cross-Architecture Generalization To verify that HF-Skip generalizes beyond SpectFormer, we apply it to GFNet (Rao et al., 2021), a pure spectral architecture without attention layers. GFNet+HF-Skip achieves 84.02% on ImageNet-100 (Table 9, Appendix) versus vanilla GFNet's 81.56% ($+2.46$ points), with the same parameter count. This shows that the routing modification is not specific to SpectFormer's hybrid attention design. HF-Skip also transfers in a frozen-backbone segmentation probe: it improves VOC mIoU by 0.0229 and Kvasir Dice by 0.0160 over SpectFormer (Appendix C). We treat these as representation-transfer measurements rather than segmentation benchmark results.

Scaling Behavior At Base scale on ImageNet-100 (77M parameters, 130 epochs), HF-SpectFormer improves top-1 accuracy from 79.46% to 81.78% and top-5 accuracy from 94.26% to 95.38%. HF-Skip also improves GFNet-Base from 83.54% to 83.82% top-1. These results show that the modification remains usable at larger model capacity on ImageNet-100.

We additionally train SpectFormer-S and HF-SpectFormer-S for 300 epochs on full ImageNet-1K. Aggregate accuracy is a practical tie: SpectFormer-S achieves 79.674% top-1 and 94.68% top-5, while HF-SpectFormer-S achieves 79.678% top-1 and 94.80% top-5, a top-1 difference of only $+0.004$ points (McNemar $p = 0.987$). This null result shows that HF-Skip is not a universal ImageNet accuracy improvement. The aggregate parity is consistent with ImageNet-1K recognition being predominantly driven by global shape: fine-scale cues may matter more for only a subset of images, so localized benefits need not produce a dataset-wide improvement. An exploratory single-seed stratification is directionally consistent with this interpretation—$+0.57$ points on images jointly high in HF content and edge density ($n = 10{,}437$, nominal $p = 0.027$), $+2.20$ on the top 1%

Table 3: Parameter-matched routing controls. Δ is relative to the corresponding SpectFormer baseline. The same ordering is largely preserved across DTD and ImageNet-100. Best score and gain in each dataset are bold.

Variant	DTD Top-1	DTD Δ	ImageNet-100 Top-1	IN-100 Δ
SpectFormer baseline	21.14 ± 0.72	—	82.67 ± 0.11	—
HF-Skip: gate LF, bypass HF	$\mathbf{30.04 \pm 0.46}$	**+8.90**	$\mathbf{83.87 \pm 0.27}$	**+1.20**
Full-spectrum bypass	27.92 ± 0.69	$+6.78$	83.44 ± 0.20	$+0.77$
Random 50% of bins bypassed	25.29 ± 0.84	$+4.15$	81.78 ± 0.60	-0.89
Inverted mask: gate HF, bypass LF	21.67 ± 0.71	$+0.53$	80.01 ± 0.17	-2.66
Generic spatial residual	21.14 ± 1.35	$+0.00$	82.28 ± 0.08	-0.39

most-HF images, and -0.44 on the lowest-HF quintile—but these subgroup results involve multiple comparisons and are not treated as confirmatory. Full details are provided in Appendix I.

6 Analysis and Ablation

6.1 Parameter-Matched Routing Controls

The band-filtered datasets establish that the gain depends on surviving frequency content. We next test whether HF-Skip merely benefits from extra parameters, easier optimization, or preserving more information in general. All controlled variants have 19,661,619 parameters; the baseline has 19,661,615, a difference of only four frozen scalar placeholders. Table 3 reports five-seed FP32 results on DTD and two-seed results on ImageNet-100 (150 epochs, $\tau = 0.5$).

The controls support a mixed but frequency-specific explanation. A full-spectrum bypass obtains $+6.78$ of HF-Skip's $+8.90$ points on DTD, showing that broad information preservation contributes substantially. However, HF-Skip still beats the full bypass by 2.12 points while bypassing *fewer* coefficients, and the same ordering holds on ImageNet-100 by 0.43 points. A generic spatial residual gives no DTD gain, the inverted mask recovers only 0.53 points, and a random half-spectrum bypass lies between the inverted and targeted variants. Thus neither parameter count nor an arbitrary residual explains the result; which coefficients receive the identity route matters.

6.2 Initialization vs. Architectural Fix

A key question is whether the accuracy gap can be resolved by changing only the initial spectral gate. We test *identity initialization* (magnitude 1, phase 0) and *unit-magnitude initialization* (magnitude 1, random phase). Identity initialization closes 65.7% of the DTD gap and 77.1% of the ImageNet-100

Table 4: Initialization ablation. A better initial gate closes part, but not all, of the gap to HF-Skip.

Configuration	DTD	ImageNet-100
Original init (random $\times$ 0.02)	22.98%	82.45%
Identity init (mag=1, phase=0)	27.77%	83.46%
Unit magnitude init	27.02%	83.28%
HF-Skip	**30.27%**	**83.76%**
Identity closes (% of gap)	65.7%	77.1%

gap, but HF-Skip remains ahead by 2.50 and 0.30 points, respectively. This result does not identify a unique Fourier-operator mechanism; it establishes the narrower and relevant point that a favorable starting gate is not equivalent to maintaining an explicit identity route throughout training. Figure 3 tracks the frequency distribution of complete block outputs during matched DTD training runs. For both SpectFormer and GFNet, HF-Skip has the highest measured DC-excluded HF fraction across the training trajectory; at convergence, identity initialization lies between HF-Skip and the original gate. This ordering is consistent with the corresponding accuracy comparison, but the activation statistic does not establish that a higher HF fraction causes the accuracy difference. We therefore treat the controlled accuracy results as the primary evidence.

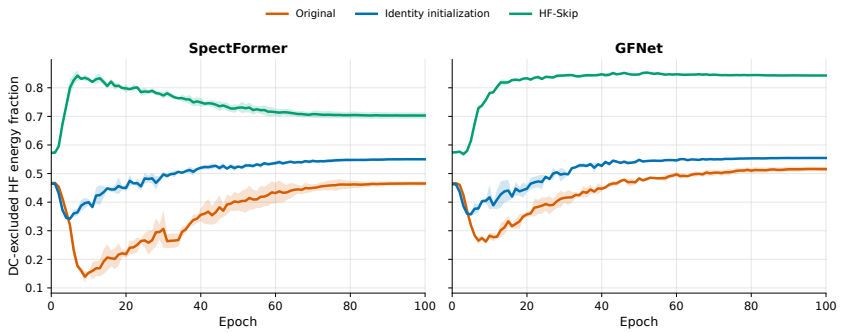

Figure 3: **Frequency distribution during DTD training.** DC-excluded HF energy fraction at the complete outputs of spectral blocks 0–3, with energies pooled across blocks, images, and channels. Curves show mean and population standard deviation over three matched seeds (42, 123, and 7), evaluated on all 1,880 validation images; epoch 0 is before training. The trajectories are descriptive and do not establish that HF fraction alone causes the accuracy differences.

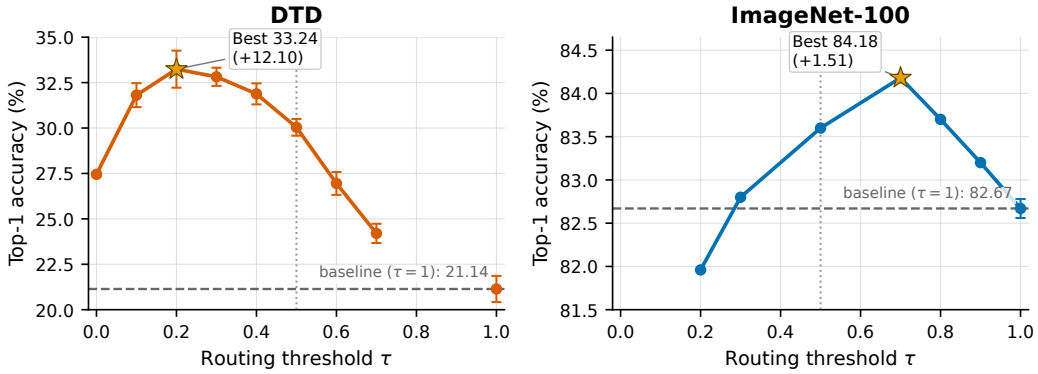

Figure 4: **Routing-threshold sweep.** Markers denote evaluated thresholds and stars identify the best Top-1 accuracy on each dataset. Horizontal dashed lines show the SpectFormer baseline at $\tau = 1$, while vertical dotted lines mark the fixed cross-dataset default $\tau = 0.5$. DTD reports mean $\pm$ standard deviation ($n = 3$; $n = 5$ at $\tau \in \{0.5, 1\}$). ImageNet-100 uses a 150-epoch schedule and is single-seed except the $\tau = 1$ anchor ($n = 2$). Panels use different vertical scales.

The important distinction is that initialization changes only the state at $t = 0$, whereas HF-Skip changes the available route at every layer and optimization step. The residual accuracy gap therefore motivates a persistent architectural intervention. **Why not regularization?** One could regularize the gate or measured band energy, but this adds a loss term and another optimization hyperparameter. HF-Skip instead changes the forward route directly; its main design choice is the threshold τ, analyzed below.

6.3 Threshold Selection

The routing threshold is dataset-dependent, so we evaluate it with a full training sweep. Figure 4 includes both endpoints: $\tau = 0$ bypasses the learned spectral gate for every bin, whereas $\tau = 1$ gates every bin and recovers the SpectFormer baseline. Exact numerical values are reported in Appendix E.

Both datasets have interior optima, but at different locations: $\tau = 0.2$ on DTD and $\tau = 0.7$ on ImageNet-100. On DTD, removing the gate entirely ($\tau = 0$) improves over baseline but remains 5.79 points below the optimum, showing that learned low-frequency processing and selective preservation are both useful. All headline cross-dataset comparisons retain the pre-specified default $\tau = 0.5$, which is optimal for neither dataset; a short validation sweep is the practical recommendation. The

Table 5: Average degradation under HF-family and LF-family corruptions. Parentheses show the change relative to the corresponding baseline.

Dataset	Model	HF-family	LF-family
ImageNet-100	SpectFormer	23.6	29.3
	HF-Skip, $\tau = 0.5$	26.0 ($+2.4$)	29.3 ($+0.0$)
	HF-Skip, $\tau = 0.7$	23.8 ($+0.2$)	28.7 (-0.6)
DTD	SpectFormer	4.0	36.2
	HF-Skip	23.6 ($+19.7$)	36.0 (-0.2)

reporting radius r used to summarize LF/HF energy is not the same quantity as the routing threshold τ and need not coincide with it.

Can τ be learned? We test a differentiable soft-mask relaxation and obtain a negative result. On DTD, learned τ moves only from 0.5 to 0.451 (per seed: 0.450, 0.451, 0.451) and matches a frozen-soft control within noise: 30.44 ± 0.57 vs. 30.51 ± 0.28 (-0.07). On ImageNet-100, the learned mean is 0.465 with a wide per-block range of 0.17–0.71, rather than converging toward the swept optimum of 0.7. In this relaxation, gradient descent does not reliably discover the validation optimum.

Other design choices. A hard and cosine-transition mask perform similarly on both datasets (Appendix F), suggesting that the result is not tied to a discontinuous cutoff. Skip-strength experiments also favor partial rather than maximal rescaling. These choices affect the magnitude of the gain, while the matched controls in Table 3 establish the more important result that the protected frequency region matters.

6.4 Band-Selective Robustness Trade-off

An identity route preserves whatever occupies its selected band: useful structure when the signal is informative, but also injected noise. We therefore group corruptions by whether they primarily perturb the HF or LF band and report accuracy degradation (lower is better) in Table 5.

The cost is confined to HF-family corruption: LF-family degradation is unchanged or slightly improved. It is also tunable. On ImageNet-100, moving from $\tau = 0.5$ to $\tau = 0.7$ reduces the HF-corruption penalty from $+2.4$ to $+0.2$ while increasing clean accuracy from 83.60 to 84.18. The DTD result is the important worst case: degradation under HF corruption increases by 19.7 points because the corruption attacks the same texture-rich band that the method preserves. HF-Skip therefore does *not* guarantee robustness to unstructured HF noise; frequency-selective preservation creates a task- and threshold-dependent accuracy–robustness trade-off. Per-corruption results remain in Appendix J. Exploratory stratified analyses show heterogeneous gains across image-detail groups on ImageNet-100 and ImageNet-1K (Appendices G and I), but remain secondary to the matched interventions; the ImageNet-1K analysis is single-seed and involves multiple comparisons.

7 Conclusion, Limitations and Future Work

We study high-frequency collapse as an architecture-dependent bottleneck in spectral vision models and propose HF-Skip, a selective identity route around the learned spectral gate. On matched ImageNet-100 data, HF-Skip gains 12.21 ± 0.25 points on HF-only inputs but only 0.67 ± 0.70 on LF-only inputs; parameter-matched controls favor the targeted route over full-spectrum, random-bin, inverted-mask, and spatial-residual alternatives. HF-Skip also yields 3-seed gains of $+1.27$ points on ImageNet-100 and $+8.83$ on DTD, transfers to GFNet, and adds only 0.009% FLOPs. Our results delineate the regimes in which HF-Skip is most useful. On full ImageNet-1K, it matches SpectFormer in aggregate accuracy, complementing the larger gains observed on controlled HF-dependent inputs and DTD. Variation in the best evaluated threshold across datasets motivates adaptive routing, while frozen-backbone probes provide initial evidence beyond classification. Future work should explore data-dependent routing, detection and high-resolution generation, and which spectral coefficients carry discriminative rather than nuisance information.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing, LaTeX preparation, and auditing and debugging analysis code, as well as to improve the grammar, phrasing, and clarity of the manuscript. All resulting code changes, calculations, figures, claims, and manuscript text were reviewed by the authors, who take full responsibility for the work.

REFERENCES

Reza Azad, Amirhossein Kazerouni, Babak Azad, Ehsan Khodapanah Aghdam, Yury Velichko, Ulas Bagci, and Dorit Merhof. Laplacian-former: Overcoming the limitations of vision transformers in local texture detection, 2023. URL https://arxiv.org/abs/2309.00108.

Jiawang Bai, Li Yuan, Shu-Tao Xia, Shuicheng Yan, Zhifeng Li, and Wei Liu. Improving vision transformers by revisiting high-frequency components. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2022.

Ronen Basri, David Jacobs, Yoni Kasten, and Shira Kritchman. The convergence rate of neural networks for learned functions of different frequencies, 2019. URL https://arxiv.org/abs/1906.00425.

Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning, 2020. URL https://arxiv.org/abs/1912.01198.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, and Andrea Vedaldi. Deep filter banks for texture recognition, description, and segmentation, 2015. URL https://arxiv.org/abs/1507.02620.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth, 2023. URL https://arxiv.org/abs/2103.03404.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2022. URL https://arxiv.org/abs/1811.12231.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive fourier neural operators: Efficient token mixers for transformers, 2022. URL https://arxiv.org/abs/2111.13587.

Jeremy Howard. Imagenette: A smaller subset of imagenet. https://github.com/fastai/imagenette, 2019.

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), 2020. ISSN 1991-7120. doi: 10.4208/cicp.oa-2020-0085. URL http://dx.doi.org/10.4208/cicp.OA-2020-0085.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms, 2022. URL https://arxiv.org/abs/2105.03824.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. URL https://arxiv.org/abs/2103.14030.

B. Ma, J. Li, Z. Jin, W. Zhang, X. Song, and B. Jin. Fine-grained image recognition with bio-inspired gradient-aware attention. *Biomimetics*, 10(12):834, December 2025. doi: 10.3390/biomimetics10120834. URL https://doi.org/10.3390/biomimetics10120834.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification, 2021. URL https://arxiv.org/abs/1905.10947.

Namuk Park and Songkuk Kim. How do vision transformers work?, 2022. URL https://arxiv.org/abs/2202.06709.

Badri N. Patro, Vinay P. Namboodiri, and Vijay Srinivas Agneeswaran. Spectformer: Frequency and attention is what you need in a vision transformer, 2023. URL https://arxiv.org/abs/2304.06446.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks?, 2022. URL https://arxiv.org/abs/2108.08810.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pp. 5301–5310. PMLR, 2019.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification, 2021. URL https://arxiv.org/abs/2107.00645.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248, June 2017. ISSN 1523-9829. doi: 10.1146/annurev-bioeng-071516-044442.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020.

Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain, 2020. URL https://arxiv.org/abs/2002.12416.

Ting Yao, Yingwei Pan, Yehao Li, Chong-Wah Ngo, and Tao Mei. Wave-vit: Unifying wavelet and transformers for visual representation learning, 2022. URL https://arxiv.org/abs/2207.04978.

Guhnoo Yun, Juhan Yoo, Kijung Kim, Jeongho Lee, and Dong Hwan Kim. Spanet: Frequency-balancing token mixer using spectral pooling aggregation modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6113–6124, 2023.

A Multi-Seed Evaluation

All experiments use three seeds (42, 123, 7) under identical training configurations.

Table 6: Multi-seed results on ImageNet-100 (150 epochs).

Seed	SpectFormer (Top-1 / Top-5)	HF-SpectFormer (Top-1 / Top-5)	Δ Top-1
42	82.48 / 95.42	83.22 / 96.00	+0.74
123	82.54 / 95.72	84.30 / 96.22	+1.76
7	82.46 / 95.50	83.76 / 95.70	+1.30
Mean $\pm$ std	**82.49 $\pm$ 0.03**	**83.76 $\pm$ 0.44**	**+1.27**

Table 7: Multi-seed results on ImageNet-100 (100 epochs).

Seed	SpectFormer (Top-1 / Top-5)	HF-SpectFormer (Top-1 / Top-5)	Δ Top-1
42	77.8 / 94.22	79.86 / 94.64	+2.06
123	78.2 / 93.88	80.18 / 94.6	+1.98
7	78.22 / 93.86	80.38 / 94.36	+2.16
Mean ± std	**78.07 ± 0.19**	**80.14 ± 0.21**	**+2.07**

Table 8: Multi-seed results on DTD (100 epochs). Values in the final row are mean ± standard deviation.

Seed	SpectFormer (Top-1 / Top-5)	HF-SpectFormer (Top-1 / Top-5)	Δ Top-1
42	23.56 / 51.70	30.21 / 62.29	+6.65
123	20.32 / 48.46	30.21 / 61.33	+9.89
7	20.43 / 47.07	30.37 / 60.96	+9.94
Mean ± std	**21.44 ± 1.51 / 49.08 ± 1.94**	**30.27 ± 0.08 / 61.52 ± 0.56**	**+8.83**

B Cross-Architecture Validation: GFNet + HF-Skip

All models trained from scratch on ImageNet-100 for 150 epochs with identical hyperparameters.

Table 9: HF-Skip applied to GFNet, a pure spectral architecture without attention layers.

Model	ImageNet-100 Top-1	Δ vs baseline
GFNet (baseline)	81.56%	—
GFNet + HF-Skip	84.02%	+2.46
SpectFormer (baseline, multi-seed mean)	82.49%	—
HF-SpectFormer (multi-seed mean)	83.76%	+1.27

B.1 Untrained Cross-Architecture Context

Figure 5 places the matched initialization result from Figure 1 alongside ViT and GFNet. The same images, five initialization seeds, and DC-excluded HF-energy metric are used for all four models. Only SpectFormer and HF-SpectFormer constitute the parameter-matched architectural comparison used in our controlled analysis. ViT and GFNet differ in architecture and parameterization and are included solely as descriptive context; cross-architecture differences in this figure are not interpreted causally.

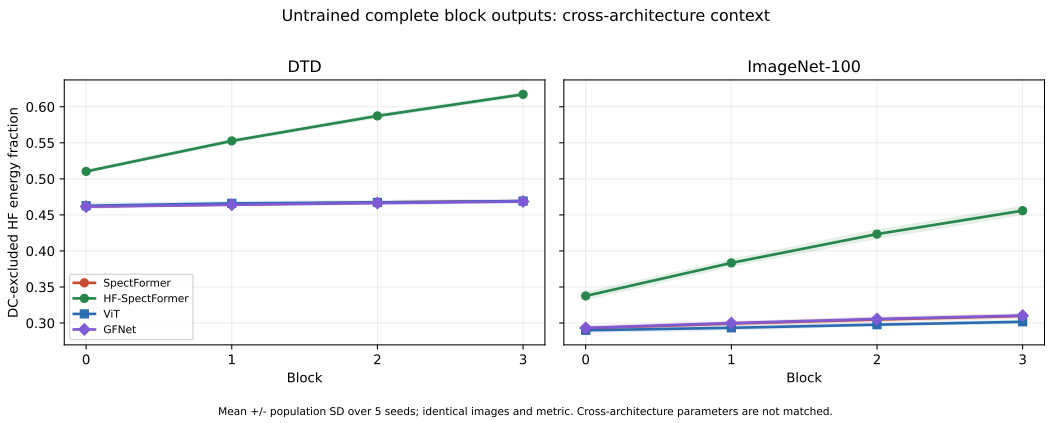

Figure 5: **Untrained complete block outputs in cross-architecture context.** DC-excluded high-frequency energy fraction at blocks 0–3 on DTD and ImageNet-100. Curves show the mean and population standard deviation over five random initializations, using identical validation images and the same frequency metric. SpectFormer and HF-SpectFormer are the matched pair; ViT and GFNet provide descriptive context only and are not parameter- or architecture-matched controls.

C Frozen-Backbone Segmentation Probe

We evaluate transfer beyond classification using frozen backbones and a lightweight decoder on two segmentation datasets. This is a controlled representation probe, not a competitive segmentation benchmark; the low absolute scores reflect the deliberately restricted decoder and training protocol.

Table 10: Frozen-backbone segmentation probe on natural (VOC 2012) and medical (Kvasir-SEG) images.

	VOC 2012		**Kvasir-SEG**	
Model	mIoU	Pixel Acc	Dice	IoU
GFNet	0.100	0.735	0.680	0.559
SpectFormer	0.119	0.746	0.723	0.619
ViT	0.136	0.753	0.753	0.643
HF-Skip	**0.142**	**0.761**	0.739	0.626

On VOC 2012, HF-Skip improves mIoU from 0.1192 to 0.1421. The smaller gain relative to DTD is plausible because region-overlap metrics are dominated by coarse shape agreement, so a boundary-localized improvement changes the score modestly. On Kvasir-SEG, HF-Skip improves Dice from 0.7225 to 0.7385, closing 53% of the gap from SpectFormer to ViT (0.7526). These results suggest that the representation transfers, but stronger decoders, end-to-end fine-tuning, and standard segmentation recipes are needed before making benchmark-level claims.

D Initialization Ablation

Table 11: Gating weight initialization strategies vs. HF-Skip.

Init Mode	DTD Test	ImageNet-100 Val
Original (random $\times$ 0.02)	22.98%	82.45%
Identity (mag=1, phase=0)	27.77%	83.46%
Unit magnitude	27.02%	83.28%
HF-Skip (multi-seed mean)	30.27%	83.76%

Table 12: Gap analysis: how much of the performance gap does identity initialization close?

Metric	DTD	ImageNet-100
Identity closes	65.7% of gap (4.79/7.29)	77.1% of gap (1.01/1.31)
HF-Skip still wins by	+2.50%	+0.30%

E ROUTING-THRESHOLD SWEEP

Table 13: Exact Top-1 accuracy values for the routing-threshold sweep in Figure 4. DTD reports mean $\pm$ standard deviation; $n = 3$ except $\tau \in \{0.5, 1\}$ where $n = 5$. ImageNet-100 is single-seed except the $\tau = 1$ anchor ($n = 2$).

τ	**DTD Top-1 (%)**	**ImageNet-100 Top-1 (%)**
0.0	27.45 ± 0.11	—
0.1	31.81 ± 0.66	—
0.2	$\mathbf{33.24 \pm 1.02}$	81.96
0.3	32.82 ± 0.50	82.80
0.4	31.88 ± 0.58	—
0.5	30.04 ± 0.46	83.60
0.6	26.95 ± 0.63	—
0.7	24.20 ± 0.53	**84.18**
0.8	—	83.70
0.9	—	83.20
1.0	21.14 ± 0.72	82.67 ± 0.11

F SMOOTH MASK ABLATION

Table 14: Mask shape ablation on DTD (100 epochs, seed 42).

Mask	Test Top-1	Δ vs Hard
Hard (paper default)	30.21%	baseline
Cosine ($w = 0.10$)	30.43%	+0.22 $\approx$ same
Gaussian narrow ($\sigma = 0.05$)	28.62%	-1.59
Gaussian medium ($\sigma = 0.10$)	26.33%	-3.88
Gaussian wide ($\sigma = 0.20$)	26.86%	-3.35

Table 15: Mask shape ablation on ImageNet-100 (100 epochs, seed 42).

Mask	Val Top-1	Δ vs Hard
Hard (paper default)	81.48%	baseline
Cosine ($w = 0.10$)	81.52%	+0.04 $\approx$ same
Gaussian ($\sigma = 0.15$)	81.60%	+0.12 $\approx$ same

G STRATIFIED ANALYSIS

Not all images rely equally on high-frequency information. To understand precisely when HF-Skip's gains arise, we stratify the ImageNet-100 validation set using three complementary criteria.

HF content quintiles. We compute the fraction of spectral energy in the outer 50% of the frequency radius for each validation image and divide into quintiles (Q1 = smoothest, Q5 = most detailed). As shown in Table 16, gains from HF-Skip are highly non-uniform: improvements are modest in low-HF

images (Q1, +0.60%) but substantially larger in medium-to-high HF regimes (Q2–Q4, +2.3% to +2.7%), indicating that HF-Skip primarily recovers performance where SpectFormer suffers from HF collapse.

Table 16: Accuracy stratified by image HF content quintile. HF-Skip gains concentrate in Q2–Q4 where structured texture is discriminative.

Quintile	HF Range	SpectFormer	HF-SpectFormer	ViT	GFNet
Q1 (Low HF)	0.000–0.004	81.30%	81.90%	77.90%	79.70%
Q2	0.004–0.008	81.80%	**84.50%**	80.10%	83.10%
Q3	0.008–0.015	83.60%	83.80%	80.10%	81.30%
Q4	0.015–0.029	84.10%	**86.40%**	82.30%	82.70%
Q5 (High HF)	0.029–0.289	80.70%	81.70%	78.10%	81.00%

Edge density terciles. Since edges naturally encode high-frequency information, we also stratify by edge density (fraction of Sobel-detected pixels) into Low, Medium, and High Detail terciles. SpectFormer degrades on edge-rich images, dropping from 82.91% (Medium) to 81.71% (High, -1.2%), whereas HF-SpectFormer remains stable ($\sim$83.1–83.9%) across strata.

Table 17: Accuracy stratified by edge density. SpectFormer degrades by -1.2% on high-detail images; HF-SpectFormer remains stable across all detail levels.

Detail Level	Samples	SpectFormer	HF-Skip	ViT	GFNet
Low Detail	1,650	82.30%	83.94%	79.82%	81.82%
Medium Detail	1,650	82.91%	83.94%	79.58%	81.70%
High Detail	1,700	81.71%	**83.12%**	79.71%	81.18%

Class-level failures. At the class level, we identify categories with the largest observed HF-Skip gains. For example, accuracy on *Kite* (Class 21) improves from 56.0% to 66.0% (+10.0%) with HF-SpectFormer, while *Axolotl* (Class 29), characterized by fine-grained textures, gains +12.0% (66.0% $\rightarrow$ 78.0%). Notably, 7 of the top 10 failure classes belong to the highest HF quartile (Q4), with mean HF magnitude 0.0248, suggesting an exploratory association with texture-rich categories.

H Model Efficiency and FLOPs

Table 18: FLOPs comparison (Small scale, 224×224 input).

Model	Params	GFLOPs	MFLOPs	Δ vs SpectFormer	Δ MFLOPs
SpectFormer	19.68M	7.5920	7591.96	baseline	—
HF-SpectFormer	19.68M	7.5926	7592.65	+0.009%	+0.69M
ViT	21.70M	8.4904	8490.44	+11.835%	+898.48M
GFNet	15.64M	5.7950	5795.00	-23.669%	-1796.96M

Table 19: Per-operation FLOPs breakdown (SpectFormer vs. HF-SpectFormer).

Operation	SpectFormer (M)	HF-SpectFormer (M)	Δ MFLOPs	Δ %
conv	115.68	115.68	0	0%
elementwise_add	1.58	1.75	+0.17	+10.88%
elementwise_mul	2.02	2.53	+0.52	+25.60%
fft_irfft2	9.96	9.96	0	0%
fft_rfft2	18.65	18.65	0	0%
gelu	28.90	28.90	0	0%
layernorm	9.41	9.41	0	0%
linear	7405.75	7405.75	0	0%
sigmoid	0.00	0.00	+0.00	new
Total	**7591.96**	**7592.65**	**+0.69**	**+0.009%**

All overhead arises from frequency masking and path recombination: 4 extra elementwise multiplications (for LF/HF mask application) and 4 extra additions (for path recombination), plus 4 scalar sigmoid calls.

I Full ImageNet-1K Evaluation

SpectFormer-S and HF-SpectFormer-S are trained for 300 epochs under the same recipe. Aggregate accuracy is statistically indistinguishable: 79.674 vs. 79.678 top-1 (McNemar $p = 0.987$), and 94.68 vs. 94.80 top-5. This result narrows the empirical scope of HF-Skip: on a large, predominantly shape-driven benchmark, preserving the selected band does not improve aggregate top-1 accuracy.

Table 20: Full ImageNet-1K evaluation (Small scale, 300 epochs, single seed).

Model	**Top-1**	**Top-5**	**Δ Top-1**
SpectFormer-S	79.674	94.68	—
HF-SpectFormer-S	79.678	94.80	+0.004

Exploratory per-image analysis. The paired validation predictions suggest heterogeneous effects. HF-Skip is $+0.57$ points better on the subset jointly high in HF content and edge density ($n = 10{,}437$, nominal $p = 0.027$), $+2.20$ on the top 1% most-HF images, and -0.44 on the lowest-HF quintile. These slices were examined after the aggregate result, use a single training seed, and involve multiple comparisons; accordingly, we treat the sign pattern as hypothesis-generating rather than evidence of a significant subgroup effect.

J Corruption Robustness: 10 Types × 5 Severity Levels

We evaluate all models under 10 corruption types at 5 severity levels on ImageNet-100. Bold indicates best performance per row. Tables are organized by corruption category: noise, blur, and weather/digital.

J.1 NOISE CORRUPTIONS

Table 21: Gaussian noise (std dev σ).

σ	SpectFormer	HF-Skip	ViT	GFNet	GFNet+HFSkip
0.08	75.1%	**77.0%**	69.7%	74.9%	76.8%
0.12	71.2%	73.1%	64.9%	70.8%	**73.5%**
0.18	63.9%	65.1%	55.3%	64.7%	**66.7%**
0.26	51.2%	50.2%	40.5%	**53.0%**	51.0%
0.38	30.2%	25.9%	19.5%	**33.0%**	26.1%

Table 22: Shot noise (Poisson λ; lower λ = more noise).

λ	SpectFormer	HF-Skip	ViT	GFNet	GFNet+HFSkip
60	75.4%	76.3%	69.9%	74.4%	**77.3%**
25	71.0%	71.7%	63.4%	69.8%	**73.1%**
12	62.6%	64.3%	53.6%	64.3%	**65.7%**
5	46.8%	44.1%	35.0%	**50.4%**	45.6%
3	33.3%	27.9%	22.4%	**36.3%**	29.5%

Table 23: Impulse noise (salt-and-pepper probability p).

p	SpectFormer	HF-Skip	ViT	GFNet	GFNet+HFSkip
0.03	70.6%	71.0%	64.3%	70.9%	**72.0%**
0.06	62.0%	62.9%	55.3%	**65.5%**	65.3%
0.09	55.7%	55.8%	47.4%	**60.1%**	59.3%
0.17	40.3%	40.1%	33.9%	**47.4%**	46.6%
0.27	26.7%	23.8%	23.0%	**33.6%**	32.4%

Table 24: Gaussian noise robustness (extended). Top-1 accuracy (%) under increasing noise levels.

Noise Std.	SpectFormer	**HF-Skip (Ours)**	ViT	GFNet
0.00	82.30	**83.66**	79.70	81.56
0.01	57.36	**62.86**	52.18	58.78
0.02	57.40	**62.78**	52.14	58.64
0.05	58.06	**63.32**	52.20	58.86
0.10	58.18	**62.78**	52.56	59.70
0.15	58.44	**63.00**	52.92	59.80
0.20	59.02	**62.68**	52.50	60.10

J.2 BLUR CORRUPTIONS

Table 25: Defocus blur (Gaussian radius r pixels).

r	SpectFormer	HF-Skip	ViT	GFNet	GFNet+HFSkip
3	47.6%	46.3%	41.1%	**50.2%**	47.3%
4	35.6%	34.1%	30.6%	**37.1%**	33.1%
6	**21.5%**	21.0%	18.7%	**21.5%**	17.3%
8	14.9%	**15.3%**	13.0%	14.0%	10.9%
10	10.9%	**11.2%**	9.8%	9.6%	7.5%

Table 26: Motion blur (directional kernel size k pixels).

k	SpectFormer	HF-Skip	ViT	GFNet	GFNet+HFSkip
10	66.4%	66.2%	61.1%	67.1%	**67.8%**
15	54.8%	55.1%	49.8%	**58.5%**	56.6%
20	46.5%	46.4%	41.8%	**50.7%**	48.5%
25	40.0%	38.6%	35.2%	**43.3%**	40.2%
30	34.8%	33.1%	31.0%	**37.9%**	33.8%

Table 27: Zoom blur (zoom factor z).

z	SpectFormer	HF-Skip	ViT	GFNet	GFNet+HFSkip
1.11	**69.8%**	68.3%	67.6%	68.0%	69.2%
1.16	**60.8%**	58.1%	58.9%	59.6%	60.2%
1.21	**56.6%**	53.5%	55.6%	55.7%	54.9%
1.26	**50.6%**	46.1%	49.9%	50.5%	48.0%
1.31	**45.0%**	39.9%	44.6%	44.7%	42.0%

J.3 Weather and Digital Corruptions

Table 28: Snow corruption.

Parameters	SpectFormer	HF-Skip	ViT	GFNet	GFNet+HFSkip
$\mu = 0.1, \sigma = 0.3, \alpha = 0.5$	69.2%	70.2%	64.2%	70.6%	**71.7%**
$\mu = 0.2, \sigma = 0.3, \alpha = 0.5$	66.7%	68.0%	61.7%	68.7%	**70.0%**
$\mu = 0.55, \sigma = 0.3, \alpha = 0.9$	35.1%	33.1%	28.0%	**41.0%**	39.0%
$\mu = 0.55, \sigma = 0.3, \alpha = 0.85$	38.4%	36.8%	30.8%	**43.5%**	42.4%

Table 29: Frost (image blend α, frost blend β).

Parameters	SpectFormer	HF-Skip	ViT	GFNet	GFNet+HFSkip
$\alpha = 1.0, \beta = 0.4$	75.8%	78.3%	73.0%	75.7%	**78.8%**
$\alpha = 0.8, \beta = 0.6$	74.3%	76.7%	70.5%	73.5%	**77.1%**
$\alpha = 0.7, \beta = 0.7$	73.2%	75.9%	68.9%	72.5%	**76.0%**
$\alpha = 0.65, \beta = 0.7$	73.8%	76.5%	69.8%	72.9%	**76.7%**
$\alpha = 0.6, \beta = 0.75$	73.0%	76.3%	69.0%	72.2%	**75.8%**

Table 30: Brightness (offset Δ).

Δ	SpectFormer	HF-Skip	ViT	GFNet	GFNet+HFSkip
+0.1	79.2%	80.2%	76.1%	77.8%	**80.8%**
+0.2	78.0%	79.8%	75.2%	77.1%	**80.3%**
+0.3	76.7%	78.8%	73.8%	76.1%	**79.4%**
+0.4	74.7%	77.0%	70.9%	74.0%	**77.6%**
+0.5	71.8%	74.3%	68.3%	71.5%	**74.4%**

Table 31: Contrast (scale factor c; lower = less contrast).

c	SpectFormer	HF-Skip	ViT	GFNet	GFNet+HFSkip
0.40	75.6%	77.2%	71.9%	74.6%	**78.0%**
0.30	74.1%	76.0%	70.0%	73.4%	**76.1%**
0.20	70.3%	72.1%	66.6%	71.2%	**73.9%**
0.10	56.9%	59.8%	54.5%	63.3%	**67.5%**
0.05	35.5%	38.8%	32.8%	44.0%	**50.6%**

K Compute Resources

Experiments were conducted on NVIDIA H100 GPUs (80GB). The standard ImageNet-100 runs use 150 epochs, the Base-scale ImageNet-100 runs use 130 epochs, and the full ImageNet-1K comparison uses 300 epochs at Small scale. DTD band-filtered experiments use three seeds per split, the matched DTD routing controls use five seeds, and the ImageNet-100 routing controls use two seeds. We report schedules and seed counts with each table and avoid an aggregate GPU-hour estimate because the expanded experiments were executed across multiple allocations.

L DTD Training-Trajectory Measurement

To characterize how routing and initialization affect block-output spectra during optimization, we train Original, identity-initialized, and HF-Skip variants of SpectFormer-Small and GFNet-Small for 100 epochs on the official DTD split, using seeds 42, 123, and 7. At every epoch, all 1,880 validation images are evaluated in deterministic order. The reported statistic uses the complete outputs of spectral blocks 0–3, unshifted one-sided Fourier coordinates, Parseval multiplicity, exclusion of the DC coefficient, and float64 accumulation of energies before division.

At epoch 100, the pooled HF fractions for SpectFormer are 0.703 ± 0.011 (HF-Skip), 0.550 ± 0.004 (identity initialization), and 0.465 ± 0.008 (Original). For GFNet they are 0.843 ± 0.001, 0.555 ± 0.001, and 0.516 ± 0.008, respectively. Thus the converged ordering is HF-Skip, identity initialization, then Original for both architectures. These values characterize activation distributions and are not interpreted as a causal explanation of accuracy.

M FEATURE MAP VISUALIZATIONS

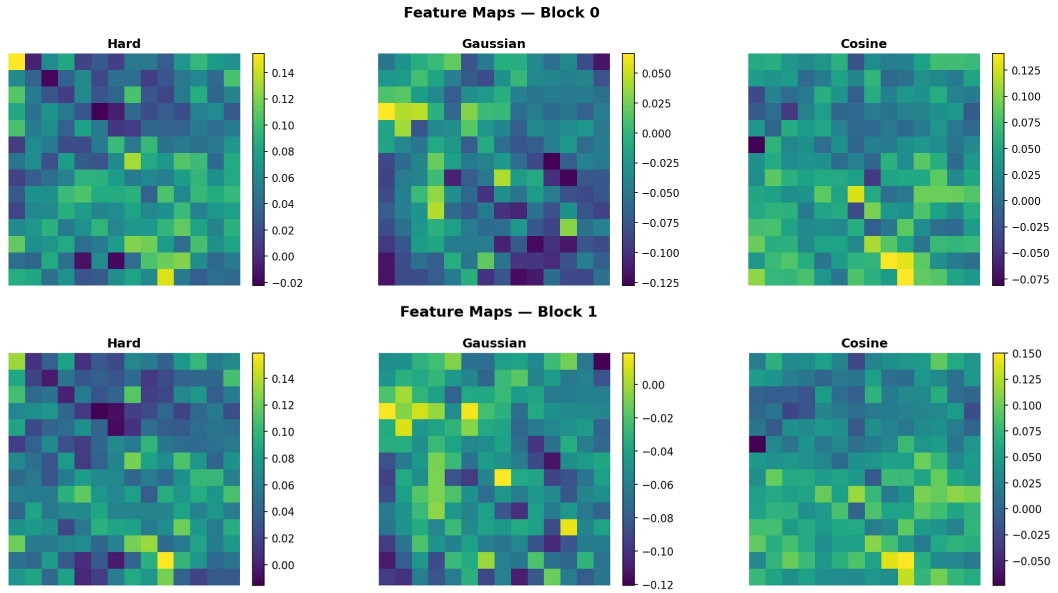

Figure 6: Feature maps for different mask types on an ImageNet-100 image.

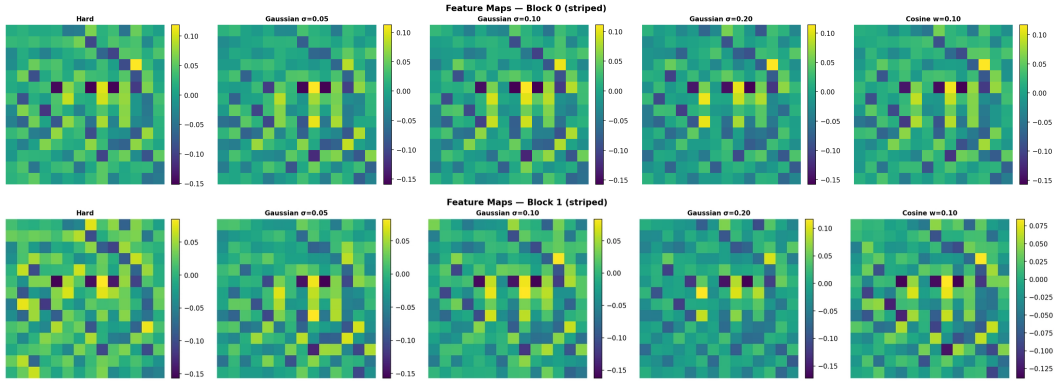

Figure 7: Feature maps for different mask types on a DTD texture image.

N MASK SHAPE VISUALIZATIONS

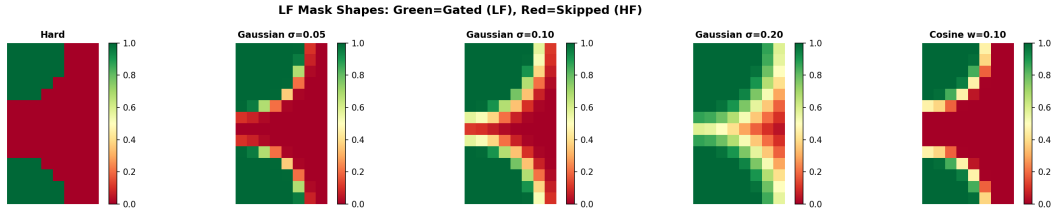

Figure 8: LF mask shapes for all five variants. Green = gated (LF), Red = skipped (HF).