FREQUENCY-AWARE PATCH WEIGHTING IN VISION TRANSFORMERS VIA THE FUNK–HECKE THEOREM

Anonymous authors
Paper under double-blind review

ABSTRACT

In Vision Transformers (ViTs), patch features are typically used directly. Existing Fourier-based methods are global and operate on the 2D image grid, where all patches share the same frequency-band weights and cannot distinguish which patches contain more useful information. As a result, the angular-frequency components of each patch (the analogue of spatial frequency in planar Fourier analysis) and the texture and detail information they carry are often ignored. In contrast, we operate on a high-dimensional unit hypersphere and weight each patch independently. We find that the spherical harmonic Fourier features of each patch allow order-indexed weighting per patch. The Gegenbauer basis functions and the Funk–Hecke theorem provide the mathematical foundation for this operation. Based on this, we propose **Funk–Hecke Adaptive Weighting** (**FHAW**), which modulates each patch individually by an order-indexed, low-order band-limited gate, giving per-patch modulation a clear angular-frequency interpretation. This mechanism opens a new path for ViT development parallel to attention. On classification, segmentation, and zero-shot recognition tasks, **FHAW** consistently outperforms the corresponding baselines with only negligible additional overhead.

1 INTRODUCTION

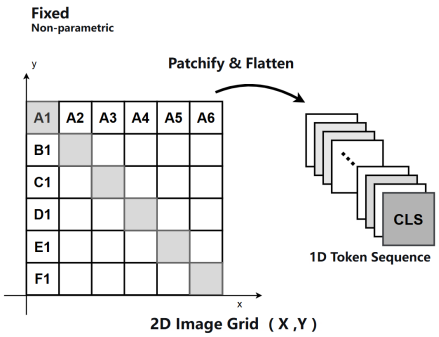
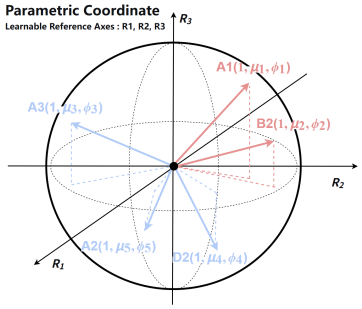
Vision Transformers (ViTs) are now a dominant backbone for image classification, semantic segmentation, and zero-shot recognition (Dosovitskiy et al., 2021; Radford et al., 2021; He et al., 2022; Oquab et al., 2024). A ViT patchifies an image, linearly projects and position-encodes patches into tokens. Self-attention mixes tokens by pairwise similarity, and the FFN transforms each token independently, so patch features serve directly as content vectors. Different patches carry different independent details. Treating them uniformly leaves these details uncontrolled. Useful details are diluted and high-frequency content fades in deeper layers, progressively weakening patch-specific information. To control such information explicitly, a natural idea is to modulate the frequency components of each patch with learnable weights. However, in Cartesian coordinates, a patch feature is represented as $x = (x_1, \ldots, x_D) \in \mathbb{R}^D$, where conventional weighting acts directly on individual feature dimensions. Such pointwise operations lack an orthogonal angular-frequency basis, making it difficult to selectively manipulate a well-defined frequency order. Arbitrary Cartesian weighting can therefore couple different angular-frequency components and obscure their spectral interpretation—a phenomenon we call the **Cartesian weighting disaster**.

In contrast, such order-specific modulation becomes naturally defined when patch features are represented in spherical coordinates. The same feature vector can be written as $(r, \theta_1, \ldots, \theta_{D-1})$, where $r = \|x\|_2$ is its magnitude and θ characterizes its angular structure. On the sphere, spherical harmonics form an orthogonal basis indexed by angular-frequency order, analogous to the Fourier basis on the 2D circle. This representation allows patch features to be decomposed into distinct angular-frequency components and each order to be modulated independently, providing a principled way to preserve, suppress, or enhance patch-specific information while leaving token mixing unchanged. This raises a key question: can we effectively manipulate individual patch features in spherical space, rather than directly in Cartesian coordinates, to achieve controllable order-wise feature modulation in ViTs?

Existing work on frequency modulation falls into three categories, none solving the problem of defining an angular-frequency band per patch in the ViT's internal feature space. Frequency-domain

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.

Table 1: **Design comparison between a standard ViT and FHAW.**

	Standard Vision Transformer (ViT)	**FHAW (ours)**
Theoretical basis	Self-attention: routing information by content similarity	Harmonic analysis on the sphere (Funk–Hecke theorem)
Patch representation	The full feature vector $x_i \in \mathbb{R}^D$: length and direction	Only the direction, as a point on a unit sphere $\hat{z}_i \in \mathbb{S}^{31}$
Shared input		
Coordinate system		
Operated object	Information routing *between* tokens (content)	A per-patch weight, set by *which way* the patch points
Parameters	$\approx$ 86M / $\approx$ 72 MB training activation, 17.4 GMACs	$\approx$ 25k / $\approx$ 67 KB training activation, 5.2 MMACs
Complexity	self-attention $O(N^2)$	$O(N)$

methods (e.g., GFNet, FNO) (Rao et al., 2021b; Li et al., 2021) perform global filtering on 2D grids or feature maps, or select/augment tokens by frequency. They typically share band weights or act on channels/token selection, and cannot define per-patch frequency bands. Spherical-harmonic and graph-based spherical networks (Cohen et al., 2018; Defferrard et al., 2020) filter external spherical signals (e.g., omnidirectional images, 3D geometry), not the internal patch-feature space of a ViT. Gating and adapters (e.g., SE, FiLM) (Hu et al., 2018; Perez et al., 2018) modulate channels, positions, or activations with few parameters, but are direction-agnostic and cannot provide well-defined angular-frequency weighting. In summary, existing methods either share band weights, operate on channels or token selection, process external spherical signals, or remain direction-agnostic. None of them defines an angular-frequency band per patch in the ViT's internal feature space or gives attention an explicit frequency-control axis — while angular convergence and high-frequency dilution in deep ViTs persist (Dong et al., 2021; Lee et al., 2025).

To address the challenge of manipulating individual patches in Cartesian coordinates, we propose **Funk–Hecke Adaptive Weighting** (**FHAW**), which weights each patch individually according to the information it carries after the spherical harmonic Fourier transform. A standard ViT represents patches in a per-channel Cartesian basis, which is not ordered by spherical-harmonic degree. Pointwise Cartesian weighting therefore cannot target a specific angular-frequency band. We instead switch to spherical coordinates and expand an axially symmetric weighting function of the angle to a reference direction, not the patch itself. Each coefficient corresponds to one angular-frequency order, and each patch evaluates this function at its own angle to obtain its weight. Thus, per-patch order-indexed modulation becomes well-defined here: the Gegenbauer coefficients index spherical-harmonic orders, and each patch evaluates the resulting low-order band-limited gate at its own angle. We do not claim strict diagonalized frequency filtering: a pointwise gate couples spherical-harmonic orders, so the Gegenbauer parameterization provides an order-indexed inductive bias rather than an exact band gain. The two designs are contrasted in Table 1.

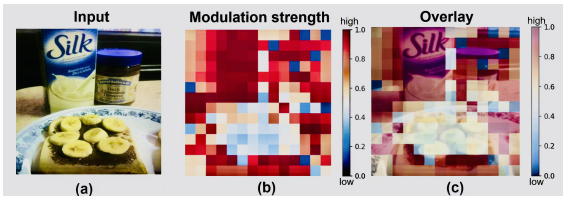


Figure 1: **Per-patch gate values g of FHAW.** Left: input; middle: g per patch (dark red: strong; blue/white: weak); right: the same g overlaid.

Our contributions are two-fold, and they describe one consistent path from theory to practice:

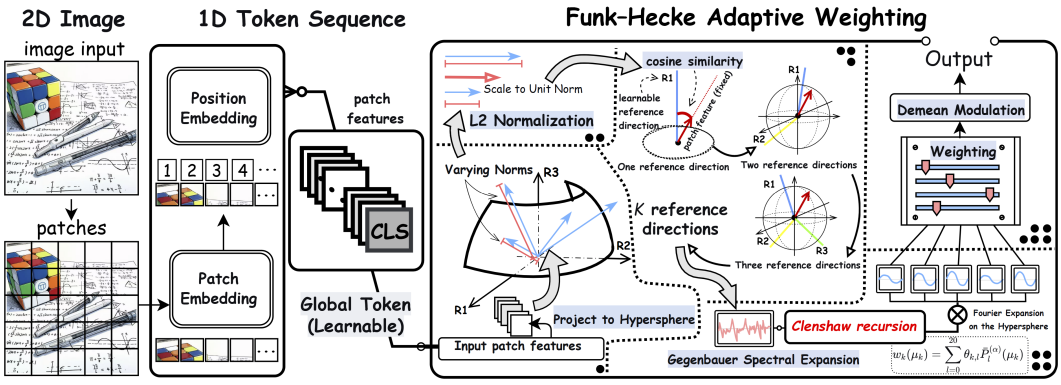


Figure 2: **The overall architecture of our method, which consists of a single technical component: the Funk–Hecke Adaptive Weighting (FHAW) module. FHAW** is inserted into the ViT pipeline after patch embedding, operating on the 1D token sequence. Note that in **FHAW**, patch features are first projected onto the unit hypersphere and then compared with K learnable reference directions via cosine similarity. A Gegenbauer spectral expansion with Clenshaw recursion computes per-patch weights, which are applied before demean modulation to produce the final output.

- **Mathematically**, we resolve the **Cartesian weighting disaster**: by moving to spherical coordinates and expanding an axially symmetric weighting function, per-patch angular-frequency weighting becomes well-defined.
- **Engineering-wise**, we build **FHAW**, a lightweight plug-and-play module with nearly four orders of magnitude fewer parameters and an order of magnitude lower complexity than the backbone. It consistently outperforms matched baselines on classification, segmentation, and zero-shot recognition, with clarified boundary conditions.

2 Related Work

Attention-based mixing. ViTs mix tokens by content similarity through self-attention, so frequency selectivity is entirely implicit and attention itself offers no explicit control over the angular-frequency content of an individual patch (Dosovitskiy et al., 2021; Liu et al., 2021; He et al., 2022; Vaswani et al., 2017). Next-ViT mixes high- and low-frequency features across its convolutional and attention blocks (Li et al., 2022), but this is an architectural-level analysis, not a learnable per-patch modulation. Deep attention also aligns token directions and dilutes high-frequency detail (Dong et al., 2021; Lee et al., 2025; Zhang et al., 2023). **FHAW** does not replace token mixing. It applies per-patch angular-frequency weighting to the already-mixed patch representations after attention.

Frequency-domain filters. Frequency-domain methods operate on 2D grids or feature maps: global filters share band weights across the image (Rao et al., 2021b; Li et al., 2021); FcaNet generalizes channel attention to the frequency domain (Qin et al., 2021); octave convolutions separate frequency components (Chen et al., 2019); frequency-aware token reduction and PHA select or augment tokens by frequency (Lee et al., 2025; Zhang et al., 2023); and scattering token mixing reweights spectral components of a fixed transform (Patro & Agneeswaran, 2023). These methods share band weights or act on channels/token selection, and cannot define an angular-frequency band per patch. **FHAW** also concerns frequency, but its object is feature direction, its coordinate system is spherical, and its granularity is per-patch.

Spherical signal processing. Spherical harmonic analysis provides the tools for frequency decomposition on the sphere (Driscoll & Healy, 1994; Atkinson & Han, 2012). Spherical CNNs exploit zonal kernels that act order by order in the spherical-harmonic basis (Cohen et al., 2018), while graph-based spherical networks build spectral filters on a sampled spherical graph (Defferrard et al., 2020); group-equivariant extensions follow the same design (Esteves et al., 2018; Kondor & Trivedi, 2018). These works filter external spherical signals—such as omnidirectional images or 3D geometry—not the internal patch-feature space learned by a ViT. **FHAW** borrows the same spherical-harmonic tools, but its object is the directional structure of patch features inside a ViT. The

underlying mathematics (the Funk–Hecke theorem, the Gegenbauer basis, spherical orthogonality) is developed in the method section.

Relation to frequency-domain methods. **FHAW** is orthogonal to existing frequency-domain methods such as GFNet, FNO, and FcaNet (Rao et al., 2021b; Li et al., 2021; Qin et al., 2021). Those methods operate on 2D grids or feature maps (GFNet, FNO) or on channels (FcaNet), whereas **FHAW** modulates the direction of each patch in the ViT's internal feature space. Because the two act on different objects, **FHAW** can be stacked on top of them. A head-to-head comparison is therefore not the appropriate evaluation. Our experiments instead isolate the gain of **FHAW** against matched baselines under identical protocols.

In summary, none of these directions defines an angular-frequency band per patch on the internal feature space of a ViT: band weights are shared, spectral operations act on channels or token selection, spherical tools target external signals, and gating remains direction-agnostic (Hu et al., 2018; Woo et al., 2018). **FHAW** closes this gap as a lightweight, plug-and-play per-patch gate that is geometry-aware, and it is complementary to attention and channel gating.

3 METHOD

3.1 SPHERICAL ANGULAR-FREQUENCY MODULATION

Spherical representation and angular frequency.

Let $x \in \mathbb{R}^d$ be the patch feature at layer l of a ViT, with $d = 768$. We apply a learnable projection

$$z = W_{\text{proj}} x, \qquad W_{\text{proj}} \in \mathbb{R}^{d_{\text{proj}} \times d}, \tag{1}$$

where $d_{\text{proj}} = 32$ determines the dimension of the sphere on which the features reside. We then apply ℓ_2 normalization:

$$\hat{z} = \frac{z}{\|z\|_2}, \qquad \|\hat{z}\|_2 = 1. \tag{2}$$

Normalization discards scale and retains only direction, so every patch lies on the unit sphere

$$\mathbb{S}^{d_{\text{proj}}-1} = \{z \in \mathbb{R}^{d_{\text{proj}}} : \|z\|_2 = 1\} = \mathbb{S}^{31}. \tag{3}$$

The similarity between two patches reduces to the inner product

$$\text{sim}(a, b) = a \cdot b = \cos\theta, \tag{4}$$

where θ is the spherical angle. Patches are compared purely by angle, and the feature cloud lies on a compact, rotationally symmetric manifold. This is the geometric substrate of our spectral machinery.

The Laplace–Beltrami operator on $L^2(\mathbb{S}^{d-1})$ has eigenpairs

$$-\Delta_{\mathbb{S}^{d-1}} \mathcal{Y}_{l,m} = \lambda_l \mathcal{Y}_{l,m}, \qquad \lambda_l = l(l + d - 2), \tag{5}$$

where $l \in \mathbb{N}_0$ is the angular-frequency order. The eigenvalue λ_l is strictly increasing in l, and the zonal component of the l-th harmonic has exactly l zeros along the polar direction (Szegő, 1975). Proofs are in App. A. Hence l can be interpreted as an angular frequency: low l corresponds to slowly varying structure, and high l to rapidly varying detail. The set $\{\mathcal{Y}_{l,m}\}$ is complete and orthonormal.

Reference directions and zonal functions.

A patch has no intrinsic orientation. Its spherical position is meaningful only relative to a reference direction. Given a reference direction $R \in \mathbb{S}^{d-1}$, a function of the form

$$h(x) = g(x \cdot R), \qquad g : [-1, 1] \to \mathbb{R}, \tag{6}$$

depends only on the angular distance $\theta = \arccos(x \cdot R)$ to R, not on the azimuth. Such functions are suitable for describing angular-position modulation relative to a learnable reference direction.

The Funk–Hecke theorem states that zonal convolution acts independently on each order, scaling the l-th order by a single scalar without mixing orders (Funk, 1916; Hecke, 1918; Stein & Weiss, 1971; Atkinson & Han, 2012). Gegenbauer polynomials are the spectral kernel that diagonalizes all zonal convolutions (Szegő, 1975; NIST, 2026). However, our module performs pointwise multiplication rather than convolution. The next paragraph makes this explicit and delineates its consequences.

Pointwise modulation and Gegenbauer spectral prior.

For each patch, the module computes its cosine similarity to a reference direction R_k

$$\mu_k = \langle \hat{z}, R_k \rangle \in [-1, 1], \tag{7}$$

and applies a learnable pointwise function

$$\tilde{p}_k = g(\mu_k) p_k. \tag{8}$$

This is pointwise multiplication, not integral convolution $\int h(x \cdot y) f(y)\, d\sigma(y)$. A pointwise multiplier couples spherical-harmonic orders: if f and g are bandlimited to L_f and L_g, the spectrum of gf generally spreads to $L_f + L_g$. Thus, **FHAW** is not a strict diagonalized frequency-domain filter, and θ_l is not the l-th-order gain of the output feature. We claim only a weaker spectral prior: the Gegenbauer parameterization provides a geometrically compatible inductive bias, with θ_l constructively associated with spherical-harmonic subspaces. $w(\mu)$ is low-order bandlimited, but the sigmoid and pointwise multiplication mix the output spectrum. L_{eff} refers to the spectral order of $w(\mu)$, not of $\tilde{\mathbf{Z}}$.

We parameterize the gate $g(\mu)$ as a truncated Gegenbauer series:

$$w(\mu) = \sum_{l=0}^{L} \theta_l \bar{P}_l^{(\alpha)}(\mu), \qquad g(\mu) = \sigma(\gamma w(\mu) + \beta). \tag{9}$$

Here L is the truncation order, controlling the degree of band limitation; θ_l is the learnable Gegenbauer coefficient of the l-th order; $\bar{P}_l^{(\alpha)}$ is the normalized Gegenbauer polynomial; α is the Gegenbauer order parameter, set to $\alpha = (d_{\text{proj}} - 2)/2$; γ, β are the learnable slope and bias of the sigmoid; and σ is the sigmoid function. Since $\bar{P}_0^{(\alpha)} \equiv 1$, the $l = 0$ term is absorbed into β. The normalized polynomials satisfy a stable three-term recurrence and are evaluated in $O(L)$ by the Clenshaw recurrence. Details are in App. B.

Why Gegenbauer: This is a prior choice, not a mathematical necessity. We make no claim that it is derived from task optimality. It is chosen for geometric and computational considerations: the gate argument satisfies $\mu \in [-1, 1]$, which is precisely the natural domain of Gegenbauer polynomials, and its orthogonality weight

$$\omega(\mu) = (1 - \mu^2)^{\alpha - 1/2} \tag{10}$$

is exactly the Jacobian of the spherical area element under $\mu = \cos\theta$, and is therefore suited to gate shapes that vary smoothly on the sphere. By the addition theorem, $\bar{P}_l^{(\alpha)}(\hat{z} \cdot R)$ is the reproducing kernel of the l-th order spherical-harmonic subspace, hence θ_l directly encodes the strength of the l-th angular-frequency component of the gate, whereas monomial coefficients do not admit such a frequency-decoupling interpretation. Under the same parameter budget, the basis further provides a geometrically adapted inductive bias and computational advantages, and is validated by ablation experiments, rather than constituting a claim of analytical optimality.

Role of the projection dimension.

Projecting to d_{proj} serves two purposes. First, it places features on the sphere $\mathbb{S}^{d_{\text{proj}}-1}$ of the spherical-representation paragraph, which makes the Gegenbauer machinery geometrically meaningful. Second, the effective parameter

$$\alpha_{\text{eff}} = \frac{d_{\text{proj}} - 2}{2} \tag{11}$$

controls how strongly higher orders are compressed over the range of μ. Larger α compresses higher orders more. For moderate α, they retain $O(1)$ magnitude and the basis remains well-conditioned. In the native space ($\alpha = 383$), higher-order basis functions are poorly conditioned over the actual μ range, whereas in the projected space ($\alpha_{\text{eff}} = 15$) they are well-conditioned (App. B). Projection is therefore a practical necessity rather than a convenience. This is an empirical, learned phenomenon rather than a fixed geometric guarantee: whether a given projected space widens or narrows the μ distribution depends on the learned W_{proj} and reference directions, and thus varies by layer. We adopt $d_{\text{proj}} = 32$. Its selection and validation are reported in the experiments.

3.2 IMPLEMENTATION

We describe a single-layer instantiation on a ViT-B/16 backbone. The module is plug-and-play, inserted between a backbone layer and the task head, with input/output shape $[B, N, D] \to [B, N, D]$, where B is batch size, $N = 196$ is the number of patches, and $D = 768$ is the feature dimension. It has about 25k trainable parameters against an 87M frozen backbone. The insertion layer is chosen in the experiments. We use $d_{\text{proj}} = 32$, $\alpha_{\text{eff}} = 15$, $K = 8$ learnable reference directions, truncation order L, and the demean modulation below. γ, β are learnable.

Projection and normalization. The filter first applies the projection $\mathbf{Z} = \mathbf{X}W_{\text{proj}}$ ($768 \to 32$) and ℓ_2 normalization:

$$\mathbf{Z} = \mathbf{X}W_{\text{proj}}, \qquad \hat{\mathbf{Z}} = \frac{\mathbf{Z}}{\|\mathbf{Z}\|_2}, \tag{12}$$

where $\mathbf{X} \in \mathbb{R}^{B\times N\times 768}$, $\mathbf{Z} \in \mathbb{R}^{B\times N\times 32}$, and each patch of $\hat{\mathbf{Z}}$ lies on $\mathbb{S}^{31}$. This step changes the conditioning of the Gegenbauer basis: measured on the trained model, the native 768-dimensional similarity has standard deviation ≈ 0.14, while the 32-dimensional projected similarity reaches ≈ 0.20, i.e., the projected space widens the μ range and keeps the retained orders well-conditioned.

Reference directions and cosine similarity. The module maintains $K = 8$ learnable reference directions

$$\{R_k\}_{k=1}^{K} \subset \mathbb{S}^{31}, \qquad \|R_k\|_2 = 1. \tag{13}$$

For every patch it computes

$$\mu_k = \langle \hat{\mathbf{Z}}, R_k \rangle \in [-1, 1], \qquad \mu = (\mu_1, \ldots, \mu_K) \in [-1, 1]^K. \tag{14}$$

Each patch is characterized by its angular position relative to the learned directions. Patches lying along a strongly activated direction receive a high μ, while orthogonal ones receive a near-zero μ.

Gegenbauer spectral response. For each reference direction, the module evaluates

$$w_k(\mu_k) = \sum_{l=0}^{L} \theta_{k,l} \bar{P}_l^{(\alpha_{\text{eff}})}(\mu_k), \qquad \alpha_{\text{eff}} = 15, \tag{15}$$

where $\theta_{k,l}$ is the learnable coefficient of the k-th reference direction and the l-th order, forming a $K \times (L + 1)$ parameter matrix. The normalized polynomials are summed via the stable Clenshaw recurrence. The zeroth-order coefficient is removed, and the constant offset is absorbed by β. Let $w_k \in \mathbb{R}^{B\times N}$ be the response for the k-th reference direction. The final scalar response is obtained by summing over the K direction responses:

$$w = \sum_{k=1}^{K} w_k \in \mathbb{R}^{B\times N}. \tag{16}$$

Summation (not averaging) lets the spectral contributions of the K directions superpose. Averaging would dilute each direction as K grows, and the removed zeroth-order coefficient leaves the scale uniformly set by γ, β.

Gate and demean modulation. The per-patch gate is

$$g = \sigma(\gamma w + \beta) \in \mathbb{R}^{B\times N}, \qquad \gamma = 5.86,\ \beta = -0.07, \tag{17}$$

where g is the per-patch soft gate and w is given above. Because g is a sigmoid, it realizes a soft selector rather than a binary switch — it multiplicatively rescales patch deviations and does not discard or replace tokens, unlike control gates used for token pruning (Rao et al., 2021a). Fig. 1 shows the resulting per-patch gate values on an example image. The colours denote the per-patch modulation strength, not a weight. The demean modulation is

$$\tilde{\mathbf{Z}} = \mathbf{Z} + g \odot (\mathbf{Z} - \bar{\mathbf{Z}}), \tag{18}$$

where

$$\bar{\mathbf{Z}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{Z}_{n,:} \in \mathbb{R}^{B\times 1\times D} \tag{19}$$

is the mean direction along the patch dimension, and $\tilde{\mathbf{Z}}$ is the modulated feature. $g \in \mathbb{R}^{B\times N}$ is broadcast along the channel dimension as $g \odot (\mathbf{Z} - \bar{\mathbf{Z}}) \in \mathbb{R}^{B\times N\times D}$. This expression preserves the feature scale while reweighting only the deviation from the mean patch direction. The gate function $w(\mu)$ has energy-weighted spectral order $L_{\text{eff}} \approx 3.1$ ($\approx 69\%$ in $l \leq 4$), consistent with the low-order profile of the pointwise-modulation paragraph. L_{eff} is the order of w, not of the output feature $\tilde{\mathbf{Z}}$.

3.3 FHAW DETECTOR (FHAW_DEC): LAYER-WISE SPECTRAL DIAGNOSTICS

To test whether the spectral prior of the pointwise-modulation paragraph holds in the trained model, we introduce **FHAW_Dec**, a diagnostic tool decoupled from **FHAW** training. **FHAW_Dec** performs spherical-harmonic decomposition on the backbone patch features $\mathbf{X}$ of any ViT layer and outputs: $\mu_{\text{mean}}, \mu_{\text{std}}$, the mean and standard deviation of the cosine similarity between patch features and the K reference directions; E_{high}, the high-frequency energy fraction; and $L_{\text{eff}_{90}}$, the order to accumulate 90% spectral energy. The decomposition target is the patch feature $\mathbf{X}$ output by the backbone, not the gate function $w(\mu)$ and not the modulated feature $\tilde{\mathbf{Z}}$. The spectral order of the gate function L_{eff} and the spectral order of the backbone feature $L_{\text{eff}_{90}}$ are two different quantities, labeled separately to avoid confusion. The layer-wise diagnostics yield two falsifiable predictions: (i) the deepest layer carries the largest μ_{mean}, hence the largest modulation space; (ii) the shallow layer sits on a plateau with directions sufficiently diverse, so modulation is redundant. The detailed layer-wise results and their multi-seed validation are reported in the experiments.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Every entry is a *matched pair*: the baseline and the +**FHAW** model share seed, initialization, data order, optimizer and schedule and differ only by the module, so Δ is the paired difference. Runs are single-seed unless stated otherwise. **FHAW** uses its default configuration (Sec. 3.2: $d_{\text{proj}} = 32$, $K = 8$, residual output) and is inserted in the L8–L12 band, which the layer-wise diagnostic selects (App. K). Table 3 (appendix) lists every line and hyperparameter, and each run fits on a single GPU. Table 2 collects the paired gains on all backbones and Fig. 3 plots them, with the first row the one non-paired entry.

Datasets. ImageNet-1K (IN-1K) (Deng et al., 2009) with ImageNet-21k (IN-21k)-pretrained trunks (Ridnik et al., 2021); CIFAR-10/100 (Krizhevsky & Hinton, 2009); Stanford Cars (Krause et al., 2013); ADE20K (Zhou et al., 2017); the four ImageNet-family sets (V2, Sketch, R, ReaL) (Recht et al., 2019; Wang et al., 2019; Hendrycks et al., 2021; Beyer et al., 2020); and 32 COCO-val images (Lin et al., 2014) for the diagnostic.

4.2 MAIN RESULTS ACROSS BACKBONES

Classification. On ViT-B/16, IN-1K (Deng et al., 2009), three-epoch fine-tuning under the official 2021 protocol (App. C) gives 82.84 against the published 81.07: $+1.76 = +1.09$ evaluation protocol plus $+1.02$ fine-tuning, not the module. The seed-42 paired gain ($+0.150$) does not replicate ($+0.013 \pm 0.152$, n=3), and its arms patch the last block, which this backbone cannot use: the classifier reads only the CLS token, so modulated patch tokens have no consumer and the module gets zero gradient (App. C). With corrected placement the module learns again, but the paired gain stays at $+0.04$ (n=2) on this plateau protocol; this line is reported as a boundary case and excluded from the claims below. On MAE ViT-B/16 and frozen ViT-B/16 (CIFAR-100, Cars) (Krizhevsky & Hinton, 2009; Krause et al., 2013), under the 5-epoch recipe (App. E, control 76.49) all five MAE pairs are positive, and the gain grows with modulation slots then saturates: one filter $+0.10$, two $+0.27$/$+0.29$, four $+0.34$/$+0.39$, eight ± 0. The gain is largest early, keeps its sign after convergence, and vanishes under a plateau protocol, i.e., **FHAW** preserves detail while the representation forms rather than moving a converged optimum (run-to-run spread $\approx \pm 0.05$). With the IN-21k (Ridnik et al., 2021) trunk frozen and only the head and **FHAW** trained (App. D), the paired gain reproduces on two seeds ($+0.27$, $+0.28$; curves separate from epoch 6), while stack-style, wider-band and shifted-band variants stay within noise. Injection-site comparison and the Cars line are in App. D.

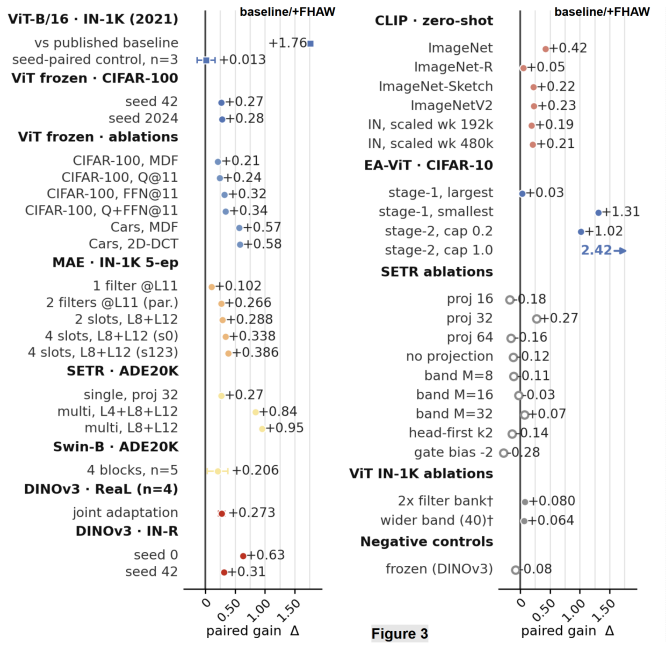


Figure 3: **Paired gains of FHAW over same-seed controls, grouped by backbone/task. Positive is better.** Rows: main results, ablations, negative controls. Values and protocols are in the appendix. †: single-seed arms of the IN-1K line that patch the inert last block (App. C).

Figure 4: **Layer-wise spectral diagnostics (FHAW_Dec, Sec. 3.3).** (a) Directional concentration μ_{mean} and (b) high-frequency energy fraction E_{high} versus block index (0–11) for six backbones: all U-shaped, peaking at block 11 (shaded band: **FHAW** insertion, 8–11). CLIP lowest.

Table 2: **Main paired results.** All Δ are paired gains in the task metric (points) over same-seed controls — except the first row, which is anchored on the published value for that backbone (see notes). Positive is better. n = number of seeds. The five-epoch MAE recipe and the CIFAR-100 variant are described in App. E and App. D.

Task / backbone	Setting	Control	+**FHAW**	Δ (n)
ViT-B/16, IN-1K	vs. published	81.07	82.84	$+1.76^{\dagger}$
MAE ViT-B/16	L8+L12	76.49	—	+0.29
ViT-B/16, frozen	CIFAR-100 base	79.47	—	+0.27 / +0.28 (2)
SETR-style, ADE20K	8/12	21.62	22.57	+0.95
Swin-B, ADE20K	4 blocks	47.34	47.55	+0.206 (5)
CLIP ViT-B/16	shift sets (four)	released weights	—	+0.05 to +0.42
DINOv3 512^2	ReaL	matched pairs	—	$+0.273 \pm 0.051$ (4)
EA-ViT, stage 1	largest → smallest	98.66 → 59.20	98.69 → 60.51	+0.03 to +1.31
EA-ViT, stage 2	budget 1.0	95.53	97.95	+2.42

Notes: the first row is anchored on the value published for this backbone (81.07, squash calibration), so its +1.76 is not the module effect: it splits into ≈+1.09 evaluation protocol and ≈+1.02 fine-tuning (App. C). †its same-run module term is $+0.013 \pm 0.152$ (n=3) and its arms patch an inert block, so the line is excluded from the claims (App. C). The SETR line is scored on a fixed 500-image val subset. DINOv3 is 4/4 positive. EA-ViT stage-2 budgets are proportional rather than absolute-compute constraints (range +0.88 to +2.42; the ten sub-models are in App. I).

Segmentation: SETR-style ViT-B/16 and Swin-B (ADE20K). (Zhou et al., 2017) A frozen IN-21k ViT-B/16 with a PUP decoder (App. F). The control is 21.62 mIoU on a fixed 500-image val subset shared by all arms. One **FHAW** layer gives +0.27, three give +0.84 and the two-layer 8/12 configuration **+0.95** (22.57) — the largest gain in this work. A 60-epoch run saturates at 21.60, so the 30-epoch result is not under-training. The projection dimension is a valley, and four auxiliary arms bound the mechanism from below (App. F). On the only hierarchical backbone here, Swin-B (Liu et al., 2021), five same-seed pairs give $\Delta = +0.206 \pm 0.175$ (5/5 positive, one-sided $p \approx 0.031$), on a baseline that sits below the released checkpoint (47.34 → 47.55; App. J).

Zero-shot: CLIP ViT-B/16. The class-text direction basis of a frozen CLIP is modulated under the official template protocol (App. G). On the four ImageNet-family sets (Recht et al., 2019; Wang et al., 2019; Hendrycks et al., 2021) the gains are $+0.42$ (ImageNet), $+0.23$ (ImageNetV2), $+0.22$ (Sketch) and $+0.05$ (ImageNet-R), each averaged over three draws of the reference weight $\mathbf{w}_k$. Benchmark datasets outside the ImageNet family are structurally out of reach for this modulation — $\mathbf{w}_k$ is a 1000-dim ImageNet class-index vector, so their classes map nowhere and the modulation closes (App. G). The direction basis is load-bearing: randomizing it degrades accuracy sharply, and single-template gains are attenuated by template ensembling (App. G).

Self-supervised: DINOv3 ViT-B/16 at 512^2. (Siméoni et al., 2025) The strongest replicated evidence in this work: $+0.273 \pm 0.051$ on ReaL (Beyer et al., 2020) over four matched seed pairs, $4/4$ positive (paired $t \approx 10.6$, one-sided $p \approx 0.0009$), about $5\times$ the within-seed noise of the metric, at no measurable cost in val (-0.005). Freezing the backbone removes the gain, and the demean output form reaches $+0.048$ (App. H).

Elastic models: EA-ViT (CIFAR-10). (Zhu et al., 2025) Ten sub-models of different width and depth come out of one adaptation run (App. I): all improve, the gain growing as capacity tightens, and stage 2 amplifies what stage 1 bakes in to $+2.42$. Control and **FHAW** share initialization, the router's Pareto front and the schedule, so the gain cannot be attributed to capacity.

4.3 Ablations: What Matters

Three patterns hold across all backbones (Table 5). (i) *The projection dimension is a valley, not monotone*: the peak is at $d_{\text{proj}} = 32$. Smaller and larger projections lose the gain. (ii) *Modulation count saturates early*: one filter is insufficient, four slots suffice, eight add nothing ($+0.10 \rightarrow +0.27 \rightarrow +0.34 \rightarrow 0$ on MAE). (iii) *The gain does not come from capacity, bandwidth, or band shift*, which leave the paired gain flat or negative ($L_{\max} = 40$, $L_{\text{init}} = 8$, stack variant: within ± 0.02 on the frozen trunk; the corpus-scale ladder is a single-seed ordering on an inert block, App. C). Two factors are load-bearing: joint optimization rather than post-hoc modulation, and the spectral parameterization itself.

The contribution of this parameterization is bounded from below by four arms on SETR and two on CLIP: a positional reference collapses to -0.65, a band form with $K = 8$ to -0.64, a head-first variant to $-0.14/-0.62$, and a forced low gate baseline to -0.28, while randomized reference directions and an independently fine-tuned head destroy the gain. The output form changes magnitude, not sign (residual $+0.273$ vs. demean $+0.048$ on DINOv3). Every value is in Table 5 (App. K).

At matched scale the training domain matters little (56.21 photos vs. 56.82 sketches), whereas $25\times$ more data adds $+20.5$ points. Decolorization costs the photo-trained models 7.4–8.7 points but the sketch model only 1.3, and grayscale inference moves the gate by $|\Delta| \leq 0.03$ while accuracy drops 7–9 points — the gate is a structural operator, not a color detector (Fig. 3; App. K). Gains require joint optimization and remaining directional headroom, and the module costs ≈ 25k parameters per filter with $O(N)$ scaling (Sec. 3.2; App. J).

5 Conclusion

We **introduce Funk–Hecke Adaptive Weighting** (**FHAW**), a lightweight module that applies per-patch angular-frequency modulation to vision transformer token representations by projecting patch features onto the unit hypersphere and constructing a band-limited Gegenbauer gate from learnable reference directions. Under joint optimization, it yields consistent paired gains across classification, segmentation, transfer, self-supervised learning, and elastic settings. Experiments suggest that these gains arise from the spectral parameterization and joint optimization rather than from added capacity, bandwidth, or band shift, which supports the promise of this direction. The method is most effective when the backbone is optimized jointly and directional structure remains to be modulated. These conditions also point to natural next steps, including broader validation, larger-scale training, more diverse tasks, and integration with attention mechanisms. Overall, **FHAW** gives a lightweight, geometry-aware route to explicit angular-frequency modulation in vision transformers, and the paired gains above show that it pays off wherever directional headroom remains.

REPRODUCIBILITY STATEMENT

We provide the source code of the FHAW filter and the FHAW_Dec diagnostic tool. The released code covers the core per-patch modulation module, the Gegenbauer/Clenshaw evaluation, the reference-direction computation, and the layer-wise spectral diagnostics. All remaining components are built on publicly available ViT backbones and standard training and evaluation pipelines. The per-line configurations, hyperparameters, random seeds, and evaluation protocols are listed in Table 3 (App. C). All datasets used in this work are publicly available, and we do not introduce any new or private data. The code will be released upon acceptance. We take responsibility for the final content of this work.

AI USE STATEMENT

In this work, we used generative AI tools for a range of assistance tasks, as disclosed in the submission form. Specifically, we used generative AI tools for: (i) language editing and polishing of the manuscript; (ii) retrieval and discovery of related literature; (iii) research ideation and execution; (iv) drafting sections of the paper; and (v) assisting with mathematical derivations and proofs. We have reviewed all AI-assisted work, including text, code, and mathematical arguments. All research ideas, experimental designs, results, and claims were verified and validated by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Kendall Atkinson and Weimin Han. *Spherical Harmonics and Approximations on the Unit Sphere: An Introduction*, volume 2044 of *Lecture Notes in Mathematics*. Springer, 2012.

Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159*, 2020.

Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolutions. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3435–3444, 2019.

Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. In *International Conference on Learning Representations (ICLR)*, 2018.

Michaël Defferrard, Martino Milani, Frédéric Gusset, and Nathanaël Perraudin. Deepsphere: a graph-based spherical CNN. In *International Conference on Learning Representations (ICLR)*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning (ICML)*, pp. 2793–2803, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

J. R. Driscoll and D. M. Jr. Healy. Computing fourier transforms and convolutions on the 2-sphere. *Advances in Applied Mathematics*, 15(2):202–250, 1994.

Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning SO(3) equivariant representations with spherical CNNs. In *European Conference on Computer Vision (ECCV)*, pp. 52–68, 2018.

Paul Funk. Über eine geometrische anwendung der abelschen integralgleichung. *Mathematische Annalen*, 77(1):129–135, 1916.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Erich Hecke. Über orthogonal-invariante integralgleichungen. *Mathematische Annalen*, 78(1):398–404, 1918.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, 2018.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning (ICML)*, pp. 2747–2755, 2018.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision (ICCV) Workshops*, pp. 554–561, 2013.

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

Dong-Jae Lee, Jiwan Hur, Jaehyun Choi, Jaemyung Yu, and Junmo Kim. Frequency-aware token reduction for efficient vision transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jiashi Li, Xin Xia, Wei Li, et al. Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios. *arXiv preprint arXiv:2207.05501*, 2022.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

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. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

NIST. Digital library of mathematical functions. National Institute of Standards and Technology, Gaithersburg, MD, 2026. https://dlmf.nist.gov/ (accessed September 2026).

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Badri Narayana Patro and Vijay Srinivas Agneeswaran. Scattering vision transformer: Spectral mixing matters. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3942–3951, 2018.

Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 15560–15572, 2021b.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning (ICML)*, pp. 5389–5400. PMLR, 2019.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik. ImageNet-21K pretraining for the masses. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Elias M. Stein and Guido Weiss. *Introduction to Fourier Analysis on Euclidean Spaces*, volume 32 of *Princeton Mathematical Series*. Princeton University Press, 1971.

Gábor Szegő. *Orthogonal Polynomials*, volume 23 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 4th edition, 1975.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 5998–6008, 2017.

Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary C. Lipton. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

Guiwei Zhang, Yongfei Zhang, Tianyu Zhang, Bo Li, and Shiliang Pu. Pha: Patch-wise high-frequency augmentation for transformer-based person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6881–6890, 2021.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 633–641, 2017.

Chen Zhu, Wangbo Zhao, Huiwen Zhang, et al. EA-ViT: Efficient adaptation for elastic vision transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

A WHY THE ORDER l IS AN ANGULAR FREQUENCY

The spherical-representation paragraph calls l the *angular frequency* of the l-th spherical harmonic on the strength of two facts: the Laplace–Beltrami spectrum grows strictly with l, and the zonal part of the l-th harmonic oscillates exactly l times along the polar direction. This appendix states both facts precisely, proves the zero count, and adds two quantitative readings. Together they make "angular frequency" a statement about the eigenfunctions rather than a name borrowed by analogy.

A.1 THE SPECTRUM GROWS STRICTLY WITH l

On the unit sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$, the Laplace–Beltrami eigenproblem reads

$$-\Delta_{\mathbb{S}^{d-1}} \mathcal{Y}_{l,m} = \lambda_l \, \mathcal{Y}_{l,m}, \qquad \lambda_l = l(l+d-2), \tag{20}$$

where $l \in \mathbb{N}_0$ is the degree and m enumerates the degenerate states of that degree. Successive eigenvalues differ by

$$\lambda_{l+1} - \lambda_l = (l+1)(l+d-1) - l(l+d-2) = 2l + d - 1, \tag{21}$$

which is positive for every $l \geq 0$ whenever $d \geq 2$, hence $\lambda_0 < \lambda_1 < \lambda_2 < \cdots$ is strictly increasing and a higher degree always carries a larger eigenvalue. Strict monotonicity alone, however, does not yet justify the word *frequency*: λ_l is a squared frequency, and the ordering says nothing by itself about how rapidly the eigenfunctions vary. The zero count supplies that second half.

A.2 THE l-TH HARMONIC OSCILLATES l TIMES

Write a point of the sphere in polar coordinates (θ, φ) relative to a fixed axis, with $\theta \in (0, \pi)$ the polar angle and φ the azimuth. For $d = 3$ the harmonics separate as $Y_{l,m}(\theta, \varphi) \propto P_l^m(\cos\theta)\, e^{im\varphi}$, with P_l^m an associated Legendre function. For general d the azimuth-independent (*zonal*) component takes the same form,

$$\mathcal{Y}_{l,0}(\theta) \;\propto\; C_l^{(\alpha)}(\cos\theta), \qquad \alpha = \frac{d-2}{2}, \tag{22}$$

with $C_l^{(\alpha)}$ the Gegenbauer polynomial already used in the reference-direction paragraph, under the same $\alpha = (d-2)/2$ and the same induced weight $(1-t^2)^{\alpha - 1/2}$.

Fact (zero count). *For $\alpha > -\frac{1}{2}$ and $l \geq 1$, $C_l^{(\alpha)}$ has exactly l simple zeros in the open interval $(-1, 1)$ (Szegő, 1975; NIST, 2026).*

Proof. Let $x_1 < \cdots < x_m$ be the points of $(-1, 1)$ at which $C_l^{(\alpha)}$ changes sign, and suppose for contradiction that $m < l$. The polynomial $q(t) = \prod_{i=1}^{m} (t - x_i)$ has degree $m < l$ and exactly the same sign changes as $C_l^{(\alpha)}$, hence the product $q(t)\, C_l^{(\alpha)}(t)$ changes sign nowhere on $(-1, 1)$ and its sign is constant there. Gegenbauer polynomials are orthogonal with respect to the weight $(1-t^2)^{\alpha-1/2}$ to every polynomial of lower degree, so

$$\int_{-1}^{1} q(t)\, C_l^{(\alpha)}(t)\, (1-t^2)^{\alpha - 1/2}\, dt = 0. \tag{23}$$

The integrand has constant sign on $(-1, 1)$ and is not identically zero, while the weight is strictly positive there, so the integral cannot vanish—a contradiction. Therefore $m \geq l$. Since $C_l^{(\alpha)}$ is a polynomial of degree l, it has at most l zeros, so $m = l$ and all of them are simple. □

Substituting $t = \cos\theta$ turns this polynomial statement into a geometric one: as θ runs from 0 to π, $\cos\theta$ decreases monotonically from 1 to -1, so the l zeros inside $(-1, 1)$ become l zeros of $\mathcal{Y}_{l,0}$ in the open interval $\theta \in (0, \pi)$, i.e. l nodal circles between the two poles. Along the polar direction the zonal component therefore oscillates exactly l times.

One precision. The exact count is a statement about the *zonal* component. For $m \neq 0$ the restriction of $\mathcal{Y}_{l,m}$ to a great circle is a trigonometric polynomial of degree at most l and hence oscillates at most l times. We state the exact count for the zonal case because that is the case our module uses (the pointwise-modulation paragraph).

A.3 Two Quantitative Readings

(a) Gradient energy (Rayleigh quotient). Since $-\Delta_{\mathbb{S}^{d-1}}\mathcal{Y}_{l,m} = \lambda_l \mathcal{Y}_{l,m}$ and the Laplace–Beltrami operator is the divergence of the surface gradient, integration by parts gives

$$\frac{\int_{\mathbb{S}^{d-1}} |\nabla \mathcal{Y}_{l,m}|^2 \, d\sigma}{\int_{\mathbb{S}^{d-1}} |\mathcal{Y}_{l,m}|^2 \, d\sigma} = \lambda_l = l(l+d-2). \tag{24}$$

Per unit L^2 energy, a degree-l function costs $l(l+d-2)$ units of variation energy: the larger l, the faster the function must change across the sphere. This is the variational counterpart of the zero count above.

(b) Asymptotic frequency. Associate to λ_l the frequency $\omega_l = \sqrt{\lambda_l}$. Then

$$\frac{\omega_l}{l} = \sqrt{1 + \frac{d-2}{l}} \;\longrightarrow\; 1 \qquad (l \to \infty), \tag{25}$$

so $\omega_l \sim l$ up to a relative error that vanishes with l. Calling l itself the angular frequency is therefore not a metaphor: for high orders the degree and the oscillation frequency agree asymptotically, and the difference from the exact value is bounded by Eq. equation 25.

A.4 Consequence for the Module

The three statements combine into the reading used in the main text: a small l labels slowly varying, global structure, and a large l labels rapidly varying, local detail. Funk–Hecke (the reference-direction paragraph) makes each order an independent scalar gain, so a learned coefficient θ_ℓ is a gain on the ℓ-th angular-frequency band rather than a mixing across bands. In our setting $d = d_{\text{proj}} = 32$, hence $\alpha_{\text{eff}} = 15$. The same identity holds in the native 768-dimensional space, where $\alpha = 383$ (the projection-dimension paragraph).

B Condition-Number Analysis and Projection Dimension

B.1 Gram Matrices

In the weighted space $L^2_\omega[-1,1]$ with $\omega(\mu) = (1-\mu^2)^{\alpha-1/2}$, the monomial basis $\{\mu^l\}$ has a dense Gram matrix $G_{ij} = \langle \mu^i, \mu^j \rangle_\omega$ that is ill-conditioned (Hilbert-type), with condition number growing exponentially in L. The normalized Gegenbauer basis is orthogonal in L^2_ω, so its Gram matrix is diagonal—the identity after L^2_ω normalization. A well-conditioned orthogonal basis means the geometry of parameter space is not skewed by the basis choice and the normal equations decouple, leading to faster convergence and coefficients less sensitive to label and sampling noise.

B.2 Projection Dimension

The effective parameter is

$$\alpha_{\text{eff}} = \frac{d_{\text{proj}} - 2}{2}. \tag{26}$$

The larger α is, the more strongly higher orders are compressed for $\mu \in (-1,1)$. For moderate α, higher orders retain $O(1)$ magnitude and the basis is well-conditioned. In the native space ($\alpha = 383$), higher-order basis functions are poorly conditioned over the actual μ range. In the projected space ($\alpha_{\text{eff}} = 15$), they are well-conditioned. We select $d_{\text{proj}} = 32$ by sweeping over $\{8, 16, 32, 64\}$ and validate it on the trained model, with a peak at 32.

B.3 Normalized Gegenbauer Polynomials and Clenshaw Summation

The unnormalized $C_l^{(\alpha)}$ overflows numerically at $\alpha = 383$ (e.g., $C_{20}^{(383)}(1) \approx 2.5 \times 10^{39}$). We use the normalized form

$$\bar{P}_l^{(\alpha)}(\mu) = \frac{C_l^{(\alpha)}(\mu)}{C_l^{(\alpha)}(1)}, \qquad C_l^{(\alpha)}(1) = \frac{(2\alpha)_l}{l!}, \tag{27}$$

Table 3: **Per-line protocols and configurations.** Δ: paired gain in the task metric (points) over the same-seed control. Positive is better. Every line is paired against its own same-protocol control. Special cases and scope limits are at the end of this appendix.

Line	Backbone / init	Train → eval	Recipe	Placement	Config	Seeds	Control → +**FHAW** (Δ)
C	ViT-B/16, IN-21k init	3 ep IN-1K → full val (50,000)	batch 256; lr 3×10^{-5}; wd 0.1; LS 0.1; warmup 0.05; Resize-256+CC-224	layer 11 (inert)	default	3	$82.711 \rightarrow 82.725$ ($+0.013 \pm 0.152$)
D	ViT-B/16, IN-21k frozen	30 ep CIFAR-100 / Cars → test	batch 512; only head + **FHAW** trainable	default	default	2	79.47 ($+0.27$/$+0.28$)
E	MAE ViT-B/16	5 ep IN-1K → full val	batch 128; peak lr 10^{-3}; layer decay 0.75; mixup 0.8/cutmix 1.0	L8+L12	1/2/4/8 slots	1–2	76.49 ($+0.10 \rightarrow +0.27 \rightarrow +0.34 \rightarrow 0$)
F	ViT-B/16 + PUP, IN-21k frozen	30 ep ADE20K → 500-image val	lr 10^{-3}; wd 10^{-4}; filter lr 5×10^{-3}; filters frozen 5 ep	single / 3-layer / 8+12	default	1	$21.62 \rightarrow 22.57$ ($+0.95$)
G	CLIP ViT-B/16, frozen	zero-shot → four IN-family shift sets	80 official templates; $\lambda = 0.1$	class-text direction basis	$\approx 10^3$ values	3 draws	released weights ($+0.05$ to $+0.42$)
H	DINOv3 ViT-B/16, 512^2	20K images, 1 ep → ReaL, full val	head lr 5×10^{-5}; filter lr 2×10^{-3}; adapters lr 10^{-5}	four-filter bank	default	4	$+0.273 \pm 0.051$
I	EA-ViT, CIFAR-10	two-stage, 100 ep/stage → 10K val	official recipe; shared init, router Pareto front and schedule	after block 11 (full-width path)	default	1	$95.53 \rightarrow 97.95$ (stage 2: $+0.88$ to $+2.42$)
J	Swin-B + UperNet-style, ADE20K init	10K iters ADE20K → 2,000-image val	batch 4; lr 2×10^{-5}; warmup 1500; module lr 10^{-2}	blocks 4, 10, 15, 21	$L_{\max}$ 8,16	5	$47.344 \rightarrow 47.550$ ($+0.206 \pm 0.175$)

where $(2\alpha)_l = \Gamma(2\alpha + l)/\Gamma(2\alpha)$ is the Pochhammer symbol (Szegő, 1975; NIST, 2026). After normalization, $\bar{P}_l^{(\alpha)}(1) = 1$, and the polynomials satisfy the stable three-term recurrence

$$\bar{P}_0^{(\alpha)}(\mu) = 1, \qquad \bar{P}_1^{(\alpha)}(\mu) = \mu, \tag{28}$$

$$\bar{P}_{l+1}^{(\alpha)}(\mu) = A_l \mu \bar{P}_l^{(\alpha)}(\mu) - B_l \bar{P}_{l-1}^{(\alpha)}(\mu), \qquad l \geq 1, \tag{29}$$

with $O(1)$ coefficients

$$A_l = \frac{2(l+\alpha)}{2\alpha + l}, \qquad B_l = \frac{l(l+2\alpha-1)}{(2\alpha+l-1)(2\alpha+l)}, \tag{30}$$

which remain numerically stable even at $\alpha = 383$. The truncated series of the pointwise-modulation paragraph is evaluated in $O(L)$ by the Clenshaw recurrence associated with this three-term recurrence, avoiding the explicit evaluation of high-degree polynomials:

$$b_{L+1} = b_{L+2} = 0, \qquad b_l = \theta_l + A_l \mu b_{l+1} - B_{l+1} b_{l+2}, \quad l = L, \ldots, 0, \qquad w(\mu) = b_0. \tag{31}$$

C Experimental Protocols and Per-Line Configurations

Table 3 collects every experimental line: initialization, training and evaluation, the recipe, the **FHAW** placement and configuration, the number of seeds, and the paired result. The sections below keep only what the table cannot carry. Special cases and scope limits are listed at the end of this section.

Row C (ViT-B/16, ImageNet-1K). We fine-tune from the official IN-21k weights for 3 epochs and evaluate on the full 50,000-image val split. Against the *published* value (81.07, squash calibration) our final number is 82.84 (+1.76); our same weights measure 80.58 under squash calibration and 81.67 under our own (in256), so the +1.76 splits into $\approx +1.09$ evaluation protocol and $\approx +1.02$ fine-tuning. The seed-42 paired gain (+0.150) does not replicate: $+0.013 \pm 0.152$ over n=3, i.e. zero on this line. Every arm patches **layer 11**, the last block, whose modulated patch tokens have no downstream consumer (the classifier reads only the CLS token), so the module receives no gradient and the arm is a placebo. The line is excluded from the claims. On a re-run with corrected placement (layer 9, a modulation ramp, module lr 10^{-2}), the gate parameters move (versus bit-frozen layer-11 arms) but the paired gain is only +0.04 on both seeds (+0.042, +0.044), within the run-to-run spread of this line: a boundary case, not a headline number.

C.1 Special Cases and Scope Limits

- **C**: the seed-42 gain does not replicate (n=3). Every arm patches an inert block. The line is excluded from the claims.
- **F**: SETR metrics are scored on a fixed 500-image val subset shared by all arms, not on the full ADE20K validation set.
- **G**: benchmark datasets outside the ImageNet family are structurally unchanged, because $\mathbf{w}_k$ is a 1000-dim ImageNet class-index vector and classes outside ImageNet map nowhere.
- **H**: freezing the backbone removes the gain.
- **I**: single seed, single dataset. Stage-2 budgets are proportional rather than absolute-compute constraints.
- **J**: a standalone diagnostic harness, not the official recipe. Its arms sit below the released checkpoint, so the paired gain is an internal one on a below-paper baseline. Freezing the backbone removes it.

D ViT-B/16, Frozen Backbone (CIFAR-100, Cars)

Row D of Table 3: the IN-21k trunk stays frozen and only the head and **FHAW** are trained for 30 epochs (batch 512, the *cifar100_if* variant). The paired gain reproduces on two independent seeds ($+0.27$, $+0.28$, Fig. 3), the per-epoch curves separating from epoch 6. Transplanting structures that win elsewhere — a stack-style variant, a wider band ($L_{\max}=40$), a shifted band ($L_{\text{init}}=8$) — lands at 79.49, 79.47 and 79.42 against the 79.47 of the base configuration, all within noise: the gain comes from the modulation itself, not from a structural variant. On the earlier 74.78-level variant, injection-site alternatives (MDF, FFN@11, Q@11, two-site) land at $+0.21$ to $+0.34$. On Cars, MDF gives $34.95 \rightarrow 35.52$ and the 2D-DCT form 35.53.

E MAE ViT-B/16

Row E of Table 3: the official MAE pre-trained weights fine-tuned with the MAE recipe shortened to 5 epochs. Control 76.49 top-1. Every configuration improves on its control and the gain grows with the number of modulation slots and then saturates (Table 3). Two *parallel* filters at one layer ($+0.266$) match two filters at two layers ($+0.288$, 0.022 apart), so the benefit tracks the number of slots rather than depth. The gain is largest early and keeps its sign after convergence, and it vanishes on a plateau protocol (fine-tuning from an already strong 83.740 checkpoint with only the top two blocks and the head trainable, 83.617 ± 0.036 over three seeds): **FHAW** preserves detail while the representation forms rather than shifting a converged optimum. Run-to-run spread on this line is $\approx \pm 0.05$.

F SETR-Style ViT-B/16 (ADE20K)

Row F of Table 3: a frozen official IN-21k ViT-B/16 feeds a SETR-style (Zheng et al., 2021) PUP decoder trained on ADE20K. All arms are scored on one fixed 500-image val subset, so deltas are paired within this caliber while the absolute values are not full-validation numbers. One **FHAW** layer gives $+0.27$ ($21.62 \rightarrow 21.89$), three give $+0.84$ (22.46) and the two-layer 8/12 configuration $+0.95$ (22.57). The projection dimension is a valley: 21.44 at $d_{\text{proj}}=16$, 21.50 with no projection, 21.89 at 32 and 21.46 at 64, against the 21.62 control. A 60-epoch run of the single-layer setting saturates at 21.60, so the 30-epoch gain is not an under-training artifact.

G CLIP ViT-B/16 (Zero-Shot)

Row G of Table 3: **FHAW** modulates the class-text direction basis of a frozen CLIP ViT-B/16 built from the official ImageNet templates, and the filter bank holds about 10^3 trainable values. On the four ImageNet-family sets the gain is $+0.42$ (ImageNet), $+0.23$ (ImageNetV2), $+0.22$ (Sketch) and $+0.05$ (ImageNet-R), averaged over three draws of the reference weights $\mathbf{w}_k$. The draw dominates

Table 4: **EA-ViT on CIFAR-10: control → +FHAW for the ten sub-models (matched pairs, full 10K val set).** Stage 2 is read across router budgets $0.2/0.4/0.6/0.8/1.0$.

Setting	Control	+**FHAW**	Δ
stage 1, largest sub-model	98.66	98.69	$+0.03$
stage 1, 2nd	98.00	98.15	$+0.15$
stage 1, 3rd	97.02	97.03	$+0.01$
stage 1, 4th	89.31	89.48	$+0.17$
stage 1, smallest	59.20	60.51	$+1.31$
stage 2, budget 0.2	95.09	96.11	$+1.02$
stage 2, budget 0.4	95.10	96.33	$+1.23$
stage 2, budget 0.6	95.53	96.41	$+0.88$
stage 2, budget 0.8	95.41	96.66	$+1.25$
stage 2, budget 1.0	95.53	97.95	$+2.42$

at larger modulation strength, so λ stays at 0.1. The direction basis is load-bearing: randomizing the reference directions degrades accuracy sharply.

H DINOv3 ViT-B/16 at 512^2

Row H of Table 3: the head, a four-filter bank and the parameter-efficient adaptation branches of the backbone are optimized jointly over 20K images from ADE20K (Zhou et al., 2017) for one epoch. We run four seeds to completion, and metrics are ReaL and the full val set from a single 512^2 center crop. The paired gain is $+0.273 \pm 0.051$ over four matched pairs ($4/4$ positive), about $5\times$ the within-seed noise of the metric, at no measurable cost in val (-0.005). Freezing the backbone removes the gain (-0.08 ReaL, -0.54 val, -0.70 on ImageNet-R). The demean output form reaches $+0.048 \pm 0.017$ on the same pairs.

I EA-ViT (CIFAR-10)

Row I of Table 3: EA-ViT (Zhu et al., 2025) produces several sub-models of different width and depth from one adaptation run, and a router selects among them under a compute budget. **FHAW** is inserted once, after block 11 on the full-width path (identity for sub-models that skip it). Both stages are paired, control and **FHAW** sharing initialization, the same Pareto front for the router and the same schedule. The ten sub-models are in Table 4. The limits (single seed, single dataset, proportional budgets) are listed as special cases above.

J Hierarchical Backbone: Swin-B (ADE20K)

Row J of Table 3: the only hierarchical backbone in this work — Swin-B (Liu et al., 2021), windowed attention with multi-resolution stages — under a standalone diagnostic harness with an UperNet-style head (not the official mmsegmentation recipe), initialized from the released ADE20K checkpoint and fine-tuned for 10K iterations (batch 4, lr 2×10^{-5}, warmup 1500), scored single-scale (mIoU) on a 2,000-image val subset. **FHAW** modulates four blocks ($\{4, 10, 15, 21\}$, $L_{\max}=8, 16$, module lr 10^{-2}). Same-seed pairs over five seeds give $47.344 \rightarrow 47.550$, $\Delta = +0.206 \pm 0.175$ ($5/5$ positive, one-sided $p \approx 0.031$). The pairing carries the signal, since the control drifts more across seeds (SD 0.385) than the effect. Boundary: the released checkpoint reproduces its reported accuracy under this harness (47.98 vs. 47.99 single-scale, 49.63 vs. 49.57 multi-scale+flip), while every arm of this line sits below it (controls 46.8–47.7, +**FHAW** 47.2–47.8) — the gain is internal to a below-paper baseline, not an improvement over the reported number. Consistently, training only **FHAW** with the backbone frozen (≈ 0.07M parameters, 6K–12K iterations, several tap layouts) lands at -0.16 to -0.03: on the plateau there is no gain left to add.

K Layer-Wise Diagnostics and Further Ablations

Tool and caliber. FHAW_Dec (Sec. 3.3) probes one backbone layer at a time and reports its directional concentration (μ, the mean cosine between a patch feature and the mean direction) and its

Table 5: **Full ablations and layer-wise diagnostics.** All Δ are paired gains in points over same-seed controls. $L_{\max}$ = maximum order, L_{init} = initial order, M = truncation order, K = reference directions, n = seeds, SD = standard deviation. Panel B reports last-block (block 11) values. All lines concentrate directions there ($\mu = 0.75$–0.92, $E_{\text{high}} = 62$–80%, $L_{90} = 15$–21). †single-seed ordering. Every arm of that line patches the last block, which this backbone does not consume (App. C).

Panel A: ablations

Projection dim (SETR)		*Band, shift, stack (ViT frozen)*	
16	-0.18	$L_{\max} = 40$	± 0.02
32	$+0.27$	$L_{\text{init}} = 8$	± 0.02
64	-0.16	stack	± 0.02
none	-0.12	*Reference layout, gate init (SETR)*	
Modulation count (MAE)		head-first	-0.14 / -0.62
1 filter	$+0.10$	positional ref.	-0.65
2 filters	$+0.27$ / $+0.29$	bias -2	-0.28
4 filters	$+0.34$ / $+0.39$	*Band approx. (SETR)*	
8 filters	≈ 0	order 8	-0.11
Filter capacity (ViT IN-1K)†		order 16	-0.03
$25.2\text{K} \to 50.9\text{K}$	$+0.15 \to +0.08$	order 32	$+0.07$
$L_{\max} = 40$	$+0.06$	$K = 8$ band	-0.64
Output form (DINOv3)		*Modulation amount (CLIP)*	
residual	$+0.273$	$3\times$ amount	$+0.19$ / $+0.21$ (SD 0.01)
demean	$+0.048$ (n=4)	smallest	$+0.143 \pm 0.055$ (n=3)

Panel B: layer-wise diagnostics

Backbone	μ	E_{high}	L_{90}
ViT-B/16 IN-21k (frozen)	0.810	70.7%	18
MAE IN-1K	0.784	65.8%	16
DINOv3 LVD-1689M	0.919	79.6%	21
EA-ViT CIFAR-10 (full)	0.797	66.7%	16
EA-ViT pre-training	0.756	61.8%	15
CLIP ViT-B/16 (tower)	0.748	65.1%	15
SETR ADE20K (30 ep)	0.799	69.4%	17

high-order energy in the Gegenbauer basis (E_{high}; L_{90}, the order below which 90% of the spectral energy lies). We re-run all six lines under one setting: 32 COCO-val (Lin et al., 2014) images, seed 42, $d_{\text{proj}} = 32$, $L_{\max} = 24$. CLIP uses its own normalization statistics and EA-ViT runs at constraint 1.0 (the full-model row). The diagnostic is a heuristic profile, not a causal instrument: we use it to choose an insertion region, nothing more.

Findings. (i) *All six lines share one shape*: a plateau through L0–L8 (μ 0.36–0.51), a rise at L9–L10, and the largest values at L11 (μ 0.75–0.92, E_{high} 62–80%, L_{90} 15–21, Table 5). The insertion points used here (L8, L11, L12) sit in that band, where the modulation has the most directional structure to act on. (ii) *The two atypical lines are the two that gain least*: DINOv3 is U-shaped (0.91 at L0, minimum 0.46 mid-network, 0.92 at L11) and CLIP has the lowest μ throughout (0.36–0.55, the most isotropic directions) — consistent with its small frozen-backbone gains ($+0.03$ to $+0.13$). (iii) *Fine-tuning barely rewrites the spectrum*: the ADE20K-finetuned SETR backbone tracks the IN-21k backbone within 0.03 per layer (0.799 vs. 0.810 at L11), so the profile belongs to the architecture and the pre-training, not to the task.

Training-domain details. The three training sets give 76.762/28.420 (IN-1K, 1.28M photos), 56.214/15.860 (IN-1K subset, 50,889) and 22.696/56.820 (ImageNet-Sketch (Wang et al., 2019), 45,889) on the IN-1K and Sketch validation domains. The trained gate is a *structural* operator, not a color detector: its per-order coefficients are a constant of the weights,

$$g_{k,l} = \sigma\big(z_{k,l} + b_l\big), \qquad \tilde{\theta}_{k,l} = \theta_{k,l}\, g_{k,l}, \qquad L_{\text{act}} = \sum_{l=0}^{L_{\max}} \sigma\big(10\,(\bar{g}_l - 0.2)\big) \approx 6.8\text{–}7.0, \tag{32}$$

where L_{act} denotes the activated-order count, a model-level constant: probing 50,000 images yields an identical L_{act} for every image. It differs from the energy-weighted spectral order of the gate used in the Method (Sec. 3.2). Grayscale inference moves the gate by $|\Delta| \le 0.03$ in-domain while accuracy drops 7–9 points, and the sketch-trained model — the one whose gate reacts most strongly to style ($\bar{g}$ 0.663 photos vs. 0.544 sketches, Cliff's $\delta \approx 0.38$) — loses only 1.3 points. A regression on the photo domain identifies edge density as the dominant driver ($\beta_{\text{edge}} = +0.34$, $\beta_{\text{color}} = -0.16$, $R^2 \approx 0.14$). The gate tracks structure, its reaction decoupled from the accuracy loss.

Further ablations. Table 5 collects the axes. Beyond it, $L_{\max} = 40$ and $L_{\text{init}} = 8$ change nothing on the ViT frozen line, the band approximation needs $M \gtrsim 32$ to approach the continuous Clenshaw recursion, and the direction basis and the joint-optimization requirement are load-bearing (randomized directions and a frozen backbone remove the gain; the CLIP head fine-tuned independently collapses to chance). A tiny cross-layer token mixer fails in all 13 configurations we try, with a diagnostic signature (an effective order of 2–3 across its fusion layers) that the diagnostic here would flag in advance.

Cost. Two filters add $2 \times 25{,}170$ parameters (0.06% of the 86.6M-parameter model). The backbone is fine-tuned exactly as in the paired baseline, with only the filter learning rate added.