Spectral Mixture-of-Experts for Continual Learning

Supplementary Material

S1. Validating Structural Interference

S1. Theoretical Proof

Under a fixed router, the change of old-task loss decomposes exactly into self-terms and cross-terms. A cross-term is nonzero iff two conditions hold simultaneously: (i) expert co-activation on old data and (ii) non-orthogonal parameter updates under the activation-weighted geometry. Enforcing subspace orthogonality eliminates all cross-terms.

S1.1. Setup and Notation

We analyze a single LoRA-based MoE layer with a frozen backbone W_0, as defined in the main paper. For an input $x \in \mathbb{R}^d$, the layer output is:

$$f(x;\Theta) = W_0 x + \sum_{m=1}^{M} a_m(x)\Delta W_m x, \qquad (1)$$

where $\Delta W_m = A_m B_m$ is the low-rank update for expert m, and $a(x) = \mathrm{softmax}(xW_g) \in \mathbb{R}^M_{\geq 0}$ is the non-negative gating vector. To isolate interference from router drift, the router is held fixed when evaluating old tasks.

Let Θ^t be the parameters after task t. The update to expert m after task $t+1$ is:

$$\delta W_m \triangleq \Delta W_m^{t+1} - \Delta W_m^t. \qquad (2)$$

We denote the old-task data distribution as $\mathcal{D}_{\text{old}}$.

S1.2. Defining Interference

We first define the metrics that capture the geometry of expert activations on the old-task data.

Definition 1 (Activation-Weighted Kernels). We define the activation-weighted second-moment (overlap) kernels Σ_{ij} as:

$$\begin{aligned} \Sigma_{mm} &\triangleq \mathbb{E}_{x\sim\mathcal{D}_{\text{old}}}\big[a_m(x)^2\, xx^\top\big], \\ \Sigma_{ij} &\triangleq \mathbb{E}_{x\sim\mathcal{D}_{\text{old}}}\big[a_i(x)a_j(x)\, xx^\top\big], \quad i \neq j. \end{aligned} \qquad (3)$$

These kernels are positive semi-definite (PSD) as they are expectations of PSD matrices. Σ_{ij} captures the co-activation statistics of experts i and j on $\mathcal{D}_{\text{old}}$. If $a_i(x)a_j(x) \equiv 0$ (no co-activation), then $\Sigma_{ij} = 0$.

Definition 2 (Weighted Inner Product). We define a weighted inner product for any two matrices U, V using a kernel Σ:

$$\langle U, V\rangle_\Sigma \triangleq \mathrm{Tr}(U\Sigma V^\top). \qquad (4)$$

S1.3. Exact Loss Decomposition

We now quantify forgetting ΔL_{old} as the output drift. We use the squared L2 loss, which serves as a precise second-order approximation for general smooth losses under the condition that Θ^t is a local minimum for the old task, as the loss increase is then dominated by the second-order term.

$$\Delta L_{\text{old}} \triangleq \mathbb{E}_{x\sim\mathcal{D}_{\text{old}}}\left[\frac{1}{2}\left\|f(x;\Theta^{t+1}) - f(x;\Theta^t)\right\|_2^2\right]. \qquad (5)$$

With the router fixed,

$$f(x;\Theta^{t+1}) - f(x;\Theta^t) = \sum_{m=1}^{M} a_m(x)\delta W_m x. \qquad (6)$$

Substituting this back into the loss definition, we get:

$$\Delta L_{\text{old}} = \frac{1}{2}\,\mathbb{E}\left\|\sum_{m=1}^{M} a_m(x)\,\delta W_m\, x\right\|_2^2. \qquad (7)$$

We introduce the shorthand notation $v_m(x) \triangleq a_m(x)\,\delta W_m\, x$. We then use the vector norm identity $\left\|\sum_m v_m\right\|_2^2 = \sum_m \|v_m\|_2^2 + 2\sum_{i<j}\langle v_i, v_j\rangle$ to expand the squared term, which yields:

$$\Delta L_{\text{old}} = \mathbb{E}\left[\frac{1}{2}\sum_{m=1}^{M}\|v_m(x)\|_2^2 \;+\; \sum_{i<j}\langle v_i(x), v_j(x)\rangle\right]. \qquad (8)$$

Next, we leverage the linearity of expectation and trace, and rewrite the inner products using the trace identity $\langle Ax, Bx\rangle = \mathrm{Tr}(A\,xx^\top B^\top)$. This moves the expectation $\mathbb{E}$ inside each term:

$$\begin{aligned} \Delta L_{\text{old}} =& \frac{1}{2}\sum_{m=1}^{M}\mathrm{Tr}\left(\delta W_m \underbrace{\mathbb{E}\big[a_m(x)^2\, xx^\top\big]}_{\Sigma_{mm}}\delta W_m^\top\right) \\ &+ \sum_{i<j}\mathrm{Tr}\left(\delta W_i \underbrace{\mathbb{E}\big[a_i(x)a_j(x)\, xx^\top\big]}_{\Sigma_{ij}}\delta W_j^\top\right). \end{aligned} \qquad (9)$$

We define the activation-weighted inner product as $\langle U, V\rangle_\Sigma \triangleq \mathrm{Tr}(U\,\Sigma\,V^\top)$. Thus, the loss is exactly decomposed as:

$$\Delta L_{\text{old}} = \underbrace{\frac{1}{2}\sum_{m=1}^{M}\langle\delta W_m, \delta W_m\rangle_{\Sigma_{mm}}}_{\text{Self-Terms}} + \underbrace{\sum_{i<j}\langle\delta W_i, \delta W_j\rangle_{\Sigma_{ij}}}_{\text{Structural Interference}}. \qquad (10)$$

S1.4. Conditions for Structural Interference

In the decomposition (Eq. (10)), the interference term $\sum_{i<j}\langle\delta W_i, \delta W_j\rangle_{\Sigma_{ij}}$ is the source of additional forgetting. The following lemma specifies the necessary and sufficient conditions for a single interference term to be non-zero.

Lemma S1 (IFF condition for a pair). For any $i \neq j$, define the pairwise interference term

$$C_{ij} \triangleq \langle \delta W_i, \delta W_j \rangle_{\Sigma_{ij}} = \mathrm{Tr}\big(\delta W_i \, \Sigma_{ij} \, \delta W_j^\top\big). \quad (11)$$

Then the following are equivalent:

i) $C_{ij} \neq 0$ (non-zero structural interference for the pair (i,j)).

ii) (Co-activation) $\Sigma_{ij} \neq \mathbf{0}$ and (non-orthogonal updates under the activation-weighted geometry) the projections of δW_i and δW_j onto the support of Σ_{ij} are not orthogonal, i.e.,

$$\sum_{r:\, \lambda_r > 0} \lambda_r \, \big\langle \delta W_i u_r, \; \delta W_j u_r \big\rangle \neq 0, \quad (12)$$

where $\Sigma_{ij} = U \Lambda U^\top$ is the eigen-decomposition with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, $\lambda_r \geq 0$, and u_r the corresponding eigenvectors.

Proof. Since Σ_{ij} is symmetric PSD, take its eigen-decomposition $\Sigma_{ij} = U \Lambda U^\top$. Using the cyclicity of the trace and letting $\Lambda^{1/2}$ be the diagonal matrix of $\sqrt{\lambda_r}$:

$$\begin{aligned} C_{ij} &= \mathrm{Tr}\big(\delta W_i U \Lambda U^\top \delta W_j^\top\big) \\ &= \mathrm{Tr}\big((\delta W_i U \Lambda^{1/2})(\delta W_j U \Lambda^{1/2})^\top\big) \\ &= \big\langle \delta W_i U \Lambda^{1/2}, \; \delta W_j U \Lambda^{1/2} \big\rangle_F \\ &= \sum_{r=1}^{d} \lambda_r \, \big\langle \delta W_i u_r, \; \delta W_j u_r \big\rangle. \end{aligned} \quad (13)$$

This step transforms the Σ_{ij}-weighted inner product into a standard Frobenius inner product in a whitened coordinate system defined by the eigenvectors of Σ_{ij}.

If $C_{ij} \neq 0$, the weighted sum in Eq. (13) is non-zero. This requires that at least one $\lambda_r > 0$ (so $\Sigma_{ij} \neq \mathbf{0}$) and that the corresponding term $\langle \delta W_i u_r, \delta W_j u_r \rangle \neq 0$. This simultaneously gives co-activation and non-orthogonality on the support. Conversely, if $\Sigma_{ij} \neq \mathbf{0}$, then some $\lambda_r > 0$. If the updates are also non-orthogonal on this support (Eq. (12)), the sum in (Eq. (13)) is non-zero, and thus $C_{ij} \neq 0$. Equivalently, using vectorization:

$$C_{ij} = \mathrm{vec}(\delta W_i)^\top \underbrace{(I \otimes \Sigma_{ij})}_{\succeq 0} \mathrm{vec}(\delta W_j), \quad (14)$$

Thus $C_{ij} = 0$ if and only if either $\Sigma_{ij} = \mathbf{0}$ (no co-activation), or the updates are orthogonal with respect to this geometry (i.e., $\mathrm{vec}(\delta W_j)$ is in the null space of $\mathrm{vec}(\delta W_i)^\top (I \otimes \Sigma_{ij})$). □

Corollary S1.1 (Whitened features). If the old-task features are whitened at the layer, $\mathbb{E}[x x^\top] \approx I$. Then $\Sigma_{ij} = \mathbb{E}[a_i a_j \, x x^\top] \approx \mathbb{E}[a_i a_j] \, I$. In this common simplifying case:

$$C_{ij} \approx \mathbb{E}[a_i a_j] \, \langle \delta W_i, \delta W_j \rangle_F, \quad (15)$$

so pairwise interference is present iff there is co-activation ($\mathbb{E}[a_i a_j] > 0$) and the Frobenius inner product between updates is non-zero.

Corollary S1.2 (General covariance upper bound). Without whitening, the magnitude of interference is bounded by:

$$|C_{ij}| = \big|\mathrm{Tr}(\delta W_i \Sigma_{ij} \delta W_j^\top)\big| \leq \lambda_{\max}(\Sigma_{ij}) \, \|\delta W_i\|_F \, \|\delta W_j\|_F. \quad (16)$$

Thus, interference is controlled by both the co-activation energy (captured by $\lambda_{\max}(\Sigma_{ij})$) and the coherence (non-orthogonality) of the updates.

S1.5. Consequences: Necessity of Orthogonality

Lemma S1 and its corollaries prove our central claim. Assume a non-zero co-activation level on old samples ($\Sigma_{ij} \neq \mathbf{0}$). Consider minimizing the old-task drift ΔL_{old} contributed by a new expert n, under a fixed norm budget $\|\delta W_n\|_F = c > 0$ (required to fit the new task). Using the whitened approximation (Corollary S1.1) for simplicity:

$$\min_{\|\delta W_n\|_F = c} \Delta L_{\text{old}} = \text{const} + 2 \sum_{m \neq n} \mathbb{E}[a_m a_n] \, \langle \delta W_m, \delta W_n \rangle_F. \quad (17)$$

The "self-term" $\frac{1}{2}\langle \delta W_n, \delta W_n \rangle_{\Sigma_{nn}}$ is part of the constant (it only depends on the norm c, not the direction of δW_n). The minimum interference is therefore attained if and only if the cross-term sum is zero. A sufficient condition is:

$$\langle \delta W_m, \delta W_n \rangle_F = 0 \quad \text{for all } m \text{ s.t. } \mathbb{E}[a_m a_n] > 0. \quad (18)$$

Therefore, orthogonality to all previously co-activated experts is a necessary condition for minimizing structural interference. This directly motivates Spectral Experts. In our parameterization, $\Delta W_m = F_o \, (S_m \odot \Theta_m) \, F_i^H$ with disjoint masks $S_m \odot S_n = 0$. This construction enforces by design that:

$$\langle \Delta W_m, \Delta W_n \rangle_F = 0 \quad (\forall m \neq n), \quad (19)$$

As shown in the main paper, this structural orthogonality ensures all cross-terms C_{ij} vanish, eliminating structural interference by construction.

S2. Experimental Validation

We now verify experimentally that the loss decomposition in Eq. (10) indeed captures structural interference in practical MoE-based continual learning, and that the cross-terms are closely related to forgetting.

S2.1. Protocol on 11 sequential tasks

We follow the main paper and consider a CLIP ViT-B/16 model equipped with MoE LoRA adapters (22 experts, top-2 routing). The model is trained sequentially on 11 classification datasets: Aircraft, Caltech101, CIFAR100, DTD, EuroSAT, Flowers, Food, MNIST, OxfordPet, StanfordCars, and SUN397. After finishing task t, we save a checkpoint Θ_t. For every

ordered pair of tasks (old, new) with old $<$ new, we load the two checkpoints Θ_{old} and Θ_{new} and perform the following analysis: i) Freeze the router of the old task. We copy the router parameters (and task id) from Θ_{old} into Θ_{new}. Both models thus share identical gating when evaluated on the old data, so any change in old-task loss must come solely from expert updates $\{\delta W_m\}$, not from router drift. ii) Measure loss drift and its decomposition. Using the dataset $\mathcal{D}_{\text{old}}$, we run the InterferenceMeter described in the main paper. For each MoE layer it accumulates:

- The total loss drift ΔL_{old},
- The sum of self-terms

$$\Delta L_{\text{self}} = \frac{1}{2} \sum_m \langle \delta W_m, \delta W_m \rangle_{\Sigma_{mm}},$$

- The sum of cross-terms

$$\Delta L_{\text{cross}} = \sum_{i<j} \langle \delta W_i, \delta W_j \rangle_{\Sigma_{ij}}.$$

We then aggregate over all image-branch layers to obtain scalars dL_all, self_all, and cross_all. The relative structural interference used in Fig. S1(b) is defined as:

$$\rho_{\text{cross}} = \frac{\text{cross\_all}}{\text{dL\_all}} \in [0, 1].$$

iii) Measure structural-interference-induced forgetting. For the same task pair we compute:

$$F_{\text{old}\to\text{new}} = \text{Acc}_{\text{old}}(\Theta_{\text{old}}) - \text{Acc}_{\text{old}},$$

i.e., the drop in zero-shot accuracy on $\mathcal{D}_{\text{old}}$ when only the expert weights are updated and the router is held fixed. This quantity, denoted Last_drop, is plotted in Fig. S1(a). In total, we obtain 55 ordered task pairs (upper-triangular part of an 11×11 matrix).

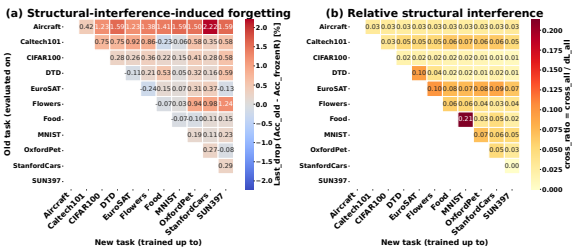

Figure S1. Empirical verification of structural interference. (a) Frozen-router forgetting ($F_{\text{old}\to\text{new}}$). Heatmap showing accuracy drops on old tasks despite a fixed router, confirming that expert parameter updates alone drive forgetting. (b) Relative interference (ρ_{cross}). Ratio of cross-term energy to total loss drift. The prevalence of non-zero cross-terms in (b) correlates strongly with the severe forgetting magnitudes in (a).

S2.2. Forgetting under a frozen router

Fig. S1(a) visualizes $F_{\text{old}\to\text{new}}$ as an upper-triangular heatmap. Several consistent patterns emerge:

- Forgetting is widespread even with a fixed router. 46 out of 55 pairs ($\approx 84\%$) exhibit positive Last_drop, i.e., old-task accuracy decreases although the router is frozen. The average drop is around 0.5% top-1, and the maximum drop is 2.22%. This shows that parameter-level interference between experts alone is sufficient to cause measurable forgetting.
- Earliest tasks are most vulnerable. The first task suffers non-trivial forgetting from almost all subsequent tasks, while later tasks such as MNIST or OxfordPet are much less affected by subsequent training. This is consistent with standard CL behavior, but here it is measured under a fixed router.
- Occasional small negative drops. A few cells are slightly blue, indicating mild positive transfer on the old task. These cases correspond to very small loss drifts and do not contradict the structural-interference view; they simply mean that the new expert updates happen to improve the old decision boundary.

Overall, Fig. S1(a) empirically confirms that the frozen-router setting isolates a non-negligible source of forgetting that cannot be explained by router drift, pointing directly to the expert update interactions analyzed in Eq. (10).

S2.3. Measuring structural interference in practice

Fig. S1(b) shows the corresponding matrix of ρ_{cross}, the relative structural interference for each task pair. We observe the following:

- Non-zero cross-terms on almost all pairs. For every pair we find $\rho_{\text{cross}} > 0$, with values ranging roughly from 0.01 to 0.21 and a mean of ≈ 0.05. Thus, in real training runs the cross-terms predicted by Eq. (10) are not a theoretical artifact: they consistently account for a noticeable fraction of the old-task loss drift.
- Strong correlation with forgetting magnitude. When we look at the absolute cross-term energy cross_all, it correlates strongly with the accuracy drop $F_{\text{old}\to\text{new}}$ across all 55 pairs (Pearson $r \approx 0.74$). Pairs that incur the largest frozen-router forgetting, such as (Aircraft, StanfordCars) and (Aircraft, SUN397), also exhibit some of the largest cross_all values. Conversely, pairs with tiny or negative Last_drop
- elative vs. absolute effect. The ratio ρ_{cross} is highest for some pairs where the total drift is small; these mostly correspond to benign cases where the structural interference component is large relative to ΔL_{old} but the absolute loss change is tiny, so the net accuracy change is close to zero. This behavior is expected from Eq. (10): cross-terms capture geometric interactions between experts, which can in principle be either harmful or mildly helpful depending on the sign.

Taken together, these observations provide direct empirical support for our theory: i) Under a fixed router,

the change in old-task loss is well explained by the decomposition into self-terms and cross-terms. ii) The cross-term energy is consistently non-zero and strongly aligned with the amount of frozen-router forgetting, confirming that structural interference is a real and measurable driver of forgetting in MoE adapters, rather than a purely conceptual notion.

S2. Proof of Proposition 1 (A Priori Orthogonality of Spectral Experts)

Proof. We use the definition of the Frobenius inner product, $\langle A, B\rangle_F = \mathrm{tr}(A^{\mathrm{H}}B)$, and the definitions of the spectral experts: $\Delta W_m = F\,(S_m \odot \Theta_m)\,F^{\mathrm{H}}$ and $\Delta W_n = F\,(S_n \odot \Theta_n)\,F^{\mathrm{H}}$. The proof proceeds as follows:

$$
\begin{aligned}
&\langle \Delta W_m, \Delta W_n\rangle_F \\
&= \mathrm{tr}\Big(\big(F(S_m\odot\Theta_m)F^{\mathrm{H}}\big)^{\mathrm{H}} \\
&\qquad \times F(S_n\odot\Theta_n)F^{\mathrm{H}}\Big) \\
&= \mathrm{tr}\Big(F(S_m\odot\Theta_m)^{\mathrm{H}}F^{\mathrm{H}}\;F(S_n\odot\Theta_n)F^{\mathrm{H}}\Big) \\
&= \mathrm{tr}\Big(F(S_m\odot\Theta_m)^{\mathrm{H}}(S_n\odot\Theta_n)F^{\mathrm{H}}\Big) \\
&= \mathrm{tr}\Big((S_m\odot\Theta_m)^{\mathrm{H}}(S_n\odot\Theta_n)\,F^{\mathrm{H}}F\Big) \\
&= \mathrm{tr}\Big((S_m\odot\Theta_m)^{\mathrm{H}}(S_n\odot\Theta_n)\Big) \\
&= \sum_{i,j}\overline{(S_m\odot\Theta_m)_{ij}}\,(S_n\odot\Theta_n)_{ij} \\
&= \sum_{i,j}\underbrace{(S_{m,ij}S_{n,ij})}_{=0}\;\overline{\Theta_{m,ij}}\;\Theta_{n,ij} \;=\; 0.
\end{aligned}
$$

The final step holds because the masks S_m and S_n are binary and pairwise disjoint. This guarantees that for any given index (i, j), the product $S_{m,ij}S_{n,ij}$ is identically zero. Therefore, every term in the summation is zero, and the entire sum vanishes regardless of the values of Θ_m and Θ_n.

S3. Derivation of Consistency Conditions

This section provides a detailed derivation of the sufficient conditions required to maintain output consistency for previously seen tasks within our proposed framework during continual learning.

S1. Consistency Objective

Let x^t be an input feature from a past task $\mathcal{T}_t$. Let $\Theta^t = \{W_g^t, \{\Theta_m^t\}_{m=1}^M\}$ denote the set of all trainable parameters after learning task $\mathcal{T}_t$. Similarly, Θ^{t+1} denotes the parameters after learning task $\mathcal{T}_{t+1}$. We define the effective weight matrix $\widetilde{W}(x;\Theta)$ for a given input x and parameter set Θ as the combination of the frozen base weights W_0 and the dynamically composed MoE spectral updates:

$$
\widetilde{W}(x;\Theta) \triangleq W_0 + \sum_{m=1}^{M} a_m(x;W_g)\Delta W_m(\Theta_m), \tag{20}
$$

where $a_m(x;W_g)$ is the gating weight for expert m produced by the router W_g, and $\Delta W_m(\Theta_m) = F(S_m \odot \Theta_m)F^H$ is the output of spectral expert m. The goal of consistency is to ensure that the model's output for the old input x^t remains unchanged after learning the new task. Using the effective weight matrix, this objective can be expressed compactly as:

$$
\widetilde{W}(x^t;\Theta^t)x^t = \widetilde{W}(x^t;\Theta^{t+1})x^t \tag{21}
$$

Expanding Eq. (21) using the definition of $\widetilde{W}$ from Eq. (20) and simplifying by cancelling the common W_0x^t term:

$$
\left(\sum_{m=1}^{M} a_m^t \Delta W_m^t\right)x^t = \left(\sum_{m=1}^{M} a_m^{t+1}\Delta W_m^{t+1}\right)x^t, \tag{22}
$$

where we use the shorthand $a^t = a(x^t;W_g^t)$, $a^{t+1} = a(x^t;W_g^{t+1})$, $\Delta W_m^t = \Delta W_m(\Theta_m^t)$, and $\Delta W_m^{t+1} = \Delta W_m(\Theta_m^{t+1})$.

S2. Router Consistency

First, we seek a condition on ΔW_g such that the gating vector remains unchanged for the old input x^t:

$$
a(x^t;W_g^t) = a(x^t;W_g^{t+1}). \tag{23}
$$

A simple sufficient condition for this equality is the equality of the inputs to the softmax function:

$$
x^tW_g^t = x^tW_g^{t+1}. \tag{24}
$$

substituting $W_g^{t+1} = W_g^t + \Delta W_g$:

$$
x^tW_g^t = x^t(W_g^t + \Delta W_g) = x^tW_g^t + x^t\Delta W_g. \tag{25}
$$

This simplifies to the router consistency constraint:

$$
x^t\Delta W_g = \mathbf{0}, \tag{26}
$$

where $\mathbf{0}$ is a zero vector of dimension $1\times M$. This implies that the router update ΔW_g must be orthogonal to the input feature x^t.

S3. Expert Composition Consistency

Assuming the router consistency condition ($a^{t+1} = a^t$) holds, the overall consistency objective (Eq. (22)) simplifies to:

$$
\left(\sum_{m=1}^{M} a_m^t \Delta W_m^t\right)x^t = \left(\sum_{m=1}^{M} a_m^t\Delta W_m^{t+1}\right)x^t, \tag{27}
$$

rearranging gives:

$$
\left(\sum_{m=1}^{M} a_m^t(\Delta W_m^{t+1} - \Delta W_m^t)\right)x^t = 0. \tag{28}
$$

Let $\Delta(\Delta W_m) = \Delta W_m^{t+1} - \Delta W_m^t$. Substituting the spectral expert definition (Eq. (3)):

$$\Delta(\Delta W_m) = F(S_m \odot \Theta_m^{t+1})F^H - F(S_m \odot \Theta_m^t)F^H. \tag{29}$$

Using the linearity of F and F^H:

$$\Delta(\Delta W_m) = F\big(S_m \odot (\Theta_m^{t+1} - \Theta_m^t)\big)F^H = F(S_m \odot \Delta\Theta_m)F^H \tag{30}$$

Substituting this back into the sum:

$$\left(\sum_{m=1}^{M} a_m^t \left[F(S_m \odot \Delta\Theta_m)F^H\right]\right) x^t = 0. \tag{31}$$

Leveraging linearity again to move F and F^H outside the sum:

$$F\left(\sum_{m=1}^{M} a_m^t (S_m \odot \Delta\Theta_m)\right) F^H x^t = 0 \tag{32}$$

Since this must hold for any past input x^t, and F is invertible, a sufficient condition is that the term within the parentheses must be the zero matrix:

$$\sum_{m=1}^{M} a_m^t (S_m \odot \Delta\Theta_m) = \mathbf{0}, \tag{33}$$

where $\mathbf{0}$ is the $d \times d$ zero matrix. Let $\Delta\Theta'_m = S_m \odot \Delta\Theta_m$ denote the sparse spectral update. The condition becomes the expert composition consistency constraint:

$$\sum_{m=1}^{M} a_m^t \Delta\Theta'_m = \mathbf{0}. \tag{34}$$

This implies the weighted sum of sparse spectral updates, using the old gating vector a^t, must be zero. Therefore, if both conditions (Eq. (26)) and (Eq. (34)) are simultaneously satisfied during the training of task $\mathcal{T}_{t+1}$, then the overall consistency objective (22) is guaranteed to hold.

S4. Metrics for Forgetting: Last FR and AUC

We consider a continual learning sequence of T tasks trained in order $1, \ldots, T$. Let $A_{t,i}$ denote the evaluation score (e.g., accuracy; any "higher-is-better" metric) on task i after finishing training on task t ($t \geq i$). Collecting these values yields an "accuracy matrix" $A \in \mathbb{R}^{T \times T}$ whose valid entries are on and above the main diagonal. Last Forgetting Rate (FR) Goal. Quantify, at the end of training, how much each past task has degraded relative to its best historical performance, and then average across tasks. For each historical task $i \in \{1, \ldots, T-1\}$,

$$B_i = \max_{t=i,\ldots,T} A_{t,i}, F_i = \frac{B_i - A_{T,i}}{B_i}. \tag{35}$$

The Last Forgetting Rate is

$$\mathrm{FR}_{\text{last}} = \frac{1}{T-1} \sum_{i=1}^{T-1} F_i \times 100\%. \tag{36}$$

Interpretation: $\mathrm{FR}_{\text{last}} = 0\%$ means no final forgetting; lower is better. This metric complements the standard Last score (final average accuracy) by measuring relative degradation from each task's own historical peak. FR Trajectory and AUC Goal. Capture forgetting throughout training, not only at the end. At each time step $t = 2, \ldots, T$, define for every past task $i \leq t-1$:

$$F_i^{(t)} = \frac{\max_{j=i,\ldots,t} A_{j,i} - A_{t,i}}{\max_{j=i,\ldots,t} A_{j,i}}. \tag{37}$$

Average across tasks to obtain the FR curve:

$$\mathrm{FR}(t) = \frac{1}{t-1} \sum_{i=1}^{t-1} F_i^{(t)} \times 100\%. \tag{38}$$

We summarize the whole curve with a normalized trapezoidal Area Under the Curve (AUC):

$$\mathrm{AUC} = \frac{1}{T-1} \sum_{t=2}^{T} \frac{\mathrm{FR}(t-1) + \mathrm{FR}(t)}{2}. \tag{39}$$

AUC is the average forgetting rate over the entire training trajectory; lower is better. Unlike $\mathrm{FR}_{\text{last}}$, AUC is sensitive to mid-training oscillations and recovery.

S5. Hyperparameter Sensitivity

S1. Analysis of Expert number N

Table S1 analyzes the number of experts N under a fixed total parameter budget. This ablation tests the "granularity" of our spectral separation, where a lower N implies fewer "larger" experts and a higher N implies more "smaller" specialized experts. The results show a clear, consistent trend: performance on all three metrics improves as N increases, with the optimal performance achieved at $N = 32$. This validates our architectural choice, suggesting our spectral framework benefits from a finer-grained, pairwise-disjoint spectral partition. Having more specialized experts enables richer, more flexible sparse compositions, enhancing both retention and generalization.

S2. Analysis of Focusing Strength γ

The focusing strength γ (Eq. (12)) controls the sharpness of the importance mapping, and Table S2shows this parameter is critical. Our chosen setting of $\gamma = 2$ achieves the optimal balance, with the highest Average score of 78.1 and Last score of 86.3. A linear mapping, where $\gamma = 1$, is too "blunt" and fails to sufficiently protect key experts, severely degrading the Last score to 76.7. Conversely, increasing the focus too much to $\gamma = 3$ or $\gamma = 4$ is also detrimental, causing both Average and Last scores to drop as the protection becomes too rigid.

N	Transfer	Average	Last
12	69.8	76.9	84.0
16	69.5	77.6	85.8
24	69.9	77.8	86.0
32	**70.1**	**78.1**	**86.3**

Table S1. Sensitivity to the number of experts N. This experiment tests different expert "granularities" (lower N = fewer, larger experts; higher N = more, smaller experts). Performance improves monotonically as N increases, reaching the optimum at $N = 32$.

Setting	Transfer	Average	Last
γ=1	67.8	72.3	76.7
γ=2	**70.1**	**78.1**	**86.3**
γ=3	69.0	74.6	80.2
γ=4	69.4	71.8	74.2

Table S2. This experiment varies γ while holding the other parameters at their optimal values ($\eta_{\text{min}} = 0.95$ and $\eta_{\text{max}} = 1.0$). $\gamma = 1$ represents a linear mapping, while $\gamma > 1$ applies a non-linear focus.

Setting	Transfer	Average	Last
η_{min}=0.00	70.2	75.3	80.4
η_{min}=0.50	69.8	77.5	85.2
η_{min}=0.95	**70.1**	**78.1**	**86.3**

Table S3. Sensitivity to the safety floor η_{min} (with $\gamma = 2$, $\eta_{\text{max}} = 1.0$). This tests the minimum stability coefficient η_m. Performance improves as the floor increases, validating our optimal choice of $\eta_{\text{min}} = 0.95$.

S3. Analysis of Safety Floor η_{min}

The safety floor η_{min} (Eq. (12)) provides a minimum level of protection to all experts, regardless of their computed importance. Table S3 clearly demonstrates this floor is essential. Setting no floor ($\eta_{\text{min}} = 0.00$) allows low-importance experts to be completely overwritten, causing a significant performance collapse; the Average score drops to 75.3 and the Last score plummets to 80.4. Performance increases monotonically with the floor, and our chosen value of $\eta_{\text{min}} = 0.95$ achieves the optimal Average score of 78.1 and the best Last score of 86.3. This confirms that a high safety floor is necessary to prevent catastrophic forgetting in less-used, but still valuable, experts.