HODGECOVER: HIGHER-ORDER TOPOLOGICAL COVERAGE DRIVES COMPRESSION OF SPARSE MIXTURE-OF-EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse Mixture-of-Experts (MoE) layers route tokens through a handful of experts, and learning-free compression of these layers reduces inference cost without retraining. An obstruction limits every existing compressor in this family: three experts can each be pairwise compatible yet form an irreducible cycle when merged together, so any score that ranks experts on pairwise signals is structurally blind to which triples are jointly mergeable. We show this cycle is a precise mathematical object, the harmonic kernel of the simplicial Laplacian on a 2-complex whose vertices are experts, whose edges carry KL merge barriers, and whose faces carry triplet barriers; Hodge-decomposing the edge-barrier signal isolates it, and triplet barriers measure joint-merge cost directly. We turn the diagnostic into a selection objective: HodgeCover greedily covers the harmonic-critical edges and triplet-critical triangles, and a hybrid variant adds weight pruning on survivors. On three open-weight MoE backbones under aggressive expert reduction, HodgeCover matches state-of-the-art learning-free baselines on the expert-reduction axis, leads on the aggressive-compression frontier of the hybrid axis, and uniquely balances retained mass across all four Hodge components. These results show that exposing the harmonic kernel changes which compressor wins under aggressive compression.

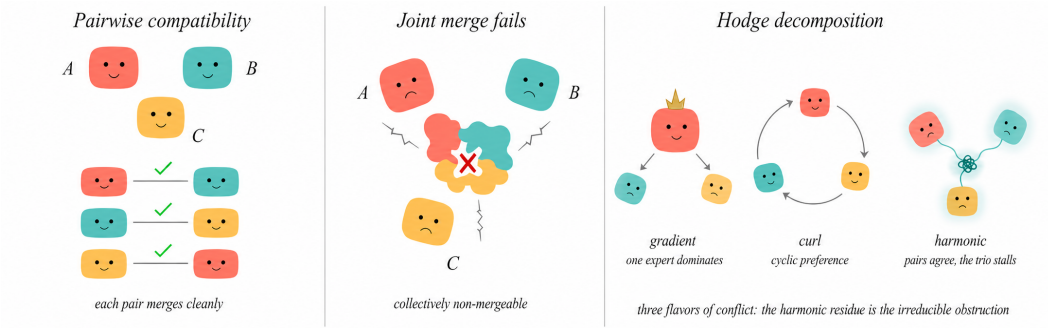

Figure 1: Three experts can be pairwise compatible yet jointly form an irreducible mergeability barrier that no pair of pairwise scores predicts. This residue is exactly the harmonic component of the Hodge decomposition on the layer's simplicial mergeability complex.

1 INTRODUCTION

In a trained sparse Mixture-of-Experts (MoE) layer, expert A may be pairwise compatible with expert B, expert B with expert C, and expert C with expert A, yet collapsing all three into a single expert can still incur an irreducible barrier that no pair of these compatibilities predicts (Fig. 1). Sparsely-gated MoE (Shazeer et al., 2017; Fedus et al., 2022) routes each token through a handful of expert sub-networks, and open checkpoints (Jiang et al., 2024; Muennighoff et al., 2024; Qwen Team, 2026; Liu et al., 2024a) host dozens to hundreds of experts per MoE layer while routing only a small subset to any one token. Most parameters therefore sit in experts that are inactive at any inference step, and a learning-free compressor removes them without fine-tuning or distillation.

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.

The dominant family of such compressors ranks experts on *pairwise* signals: REAP (Lasby et al., 2025) uses output saliency, REAM (Jha et al., 2026) reweights routing through saliency-aware merge scores, MC-SMoE (Li et al., 2023) clusters experts via routing-cosine similarity, and STUN (Lee et al., 2025) structurally prunes experts before Wanda (Sun et al., 2023) unstructured sparsity. Pairwise rankings, however, are blind to a structural property of mergeability: three experts can be pairwise mergeable yet *collectively* non-mergeable when their merge-barriers form an irreducible cycle. Methods that do consume triplet inputs, such as hypergraph spectral cuts (Zhou et al., 2006) and triplet-loss penalties (Schroff et al., 2015), aggregate triplets into a sum-of-pairs surrogate or a binary veto rather than isolating the topological residue. The harmonic kernel of the simplicial Laplacian (Lim, 2020; Schaub et al., 2020) captures this residue, and prior topological deep learning (Bronstein et al., 2021) has used it as a diagnostic tool but not as a parameter-selection objective.

To this end, we introduce *HodgeCover*, a learning-free expert-selection objective that turns the harmonic kernel into a coverage criterion. From a sparse MoE layer with n experts we build a 2-dimensional simplicial complex K whose vertices are experts, whose edges carry pairwise KL merge-barriers, and whose 2-faces carry triplet barriers, and Hodge-decompose the edge-barrier signal into gradient, curl, and harmonic components (Lim, 2020). HodgeCover then picks k survivor experts by greedy submodular coverage (Nemhauser et al., 1978) of the top-$p\%$ harmonic-critical edges and triplet-critical triangles, weighted by a saliency score adapted from REAP (Lasby et al., 2025); non-survivors are pruned with a router redirect to the closest survivor. The hybrid *HodgeCover+Wanda* adds unstructured Wanda (Sun et al., 2023) pruning on survivor weights. Our contributions are:

- We define the simplicial mergeability complex of a sparse MoE layer and identify its higher-order mergeability obstruction with the harmonic component of its Hodge decomposition, verifying this component is non-trivial ($29 - 62\%$ of per-layer barrier energy) across three MoE families (§3).
- We instantiate HodgeCover, a learning-free objective that greedily covers the top-$p\%$ harmonic-critical edges and triplet-critical triangles, and a hybrid HodgeCover+Wanda that pairs it with unstructured Wanda pruning on survivors (§4, Fig. 3).
- At 66% compression on three MoE backbones, HodgeCover+Wanda has the best perplexity on every model and leads STUN+Wanda (Lee et al., 2025) by up to $+12.6$ pp downstream accuracy; a triplet-penalty pipeline without the Hodge step trails by 5.74 pp on Qwen 3.5-35B (§5, §5.4).

2 Related Work

MoE expert pruning and merging is a static post-training family that ranks experts on pairwise signals scored from a small calibration set: REAP (Lasby et al., 2025) on output saliency, REAM (Jha et al., 2026) on saliency-aware merge scoring, MC-SMoE (Li et al., 2023) and HC-SMoE (Chen et al., 2024) on routing-cosine clustering, and STUN (Lee et al., 2025) on a structured-then-unstructured decision composed with Wanda (Sun et al., 2023) and OWL (Yin et al., 2023). Calibration-free pretrained-weight scores (Lu et al., 2024; Liu et al., 2026a) and router-gated weight metrics (Xie et al., 2024) relax the supervision of the same pairwise rule, while evolutionary search over expert subsets (Liu et al., 2024b; 2026b), prune-then-recompose redistribution (Yang et al., 2024), and unified slim-trim trimming (He et al., 2024) relax the one-shot constraint; learning-based independent-expert training and merging (Li et al., 2022; Zhong et al., 2022) and dense-to-sparse upcycling (Komatsuzaki et al., 2022) sit orthogonally on the capacity-construction axis. Every method aggregates the merge landscape into edge-level pairwise scores, and none represents the higher-order obstruction that arises when three experts are pairwise mergeable but not jointly mergeable.

Sparsity-based weight pruning compresses LLMs at the weight-matrix level along three axes. One-shot unstructured pruning ranks weights from calibration activations (Wanda (Sun et al., 2023), SparseGPT (Frantar & Alistarh, 2023)) and extends a magnitude- and movement-pruning lineage (Han et al., 2015; Sanh et al., 2020); structured pruning removes coupled channels, heads, or transformer blocks (LLM-Pruner (Ma et al., 2023), ShortGPT (Men et al., 2025)); post-training quantization compresses bit-width rather than count (GPTQ (Frantar et al., 2022), AWQ (Lin et al., 2024), SmoothQuant (Xiao et al., 2023), OmniQuant (Shao et al., 2023)). These axes operate on weight matrices and do not reduce the expert count. HodgeCover+Wanda composes the two: Stage-1 expert reduction, Stage-2 unstructured Wanda on survivor weights.

Topological and spectral methods on neural networks apply algebraic-topology and spectral machinery along two complementary lines. The architectural line installs Hodge and simplicial

operators as message-passing primitives or feature layers (geometric deep learning (Bronstein et al., 2021), simplicial neural networks (Ebli et al., 2020; Bodnar et al., 2021), topological GNNs (Horn et al., 2021)), with foundational Hodge theory traced to Eckmann (1944) and modernized for discrete data (Lim, 2020; Schaub et al., 2020; Zheng et al., 2026). The diagnostic line summarises Betti numbers, persistence diagrams, and Hodge spectra of trained networks for generalization analysis (Naitzat et al., 2020; Rieck et al., 2018; Grande & Schaub, 2024; Zheng et al., 2026); the closest matrix-level precedents that act on a spectral signal, spectral sparsification (Spielman & Teng, 2011) and spectral pruning (Suzuki et al., 2018), operate at the dyadic level and never lift to higher-order simplicial structure. Across both lines topology is consumed as input or summarised as output, not used as a parameter-selection objective.

Triplet- and hypergraph-aware methods in compression and clustering encode higher-order relational structure as hyperedges or triplet constraints in two distinct families. Hypergraph spectral methods reduce triplet (or larger) cliques to a clique-expansion Laplacian for cuts (Zhou et al., 2006; Agarwal et al., 2006), higher-order Cheeger inequalities (Lee et al., 2014; Louis, 2015), motif-based partitioning (Benson et al., 2016), and hypergraph convolution and attention (Feng et al., 2019; Bai et al., 2021). Metric learning encodes triplets as scalar margin constraints in pair-contrastive form (Hadsell et al., 2006) or triplet form (FaceNet (Schroff et al., 2015), deep ranking (Wang et al., 2014)). All these methods aggregate higher-order inputs into a sum-of-pairs surrogate or a binary veto, never separating the gradient, curl, and harmonic components of the signal; §5.4 compares such a soft-triplet penalty with HodgeCover on the same triplet inputs.

3 The Mergeability Complex and Its Hodge Decomposition

We now formalize the higher-order obstruction sketched in Section 1; the full Hodge primer, all proofs, projection formulas, and extended diagnostics are in App. B.

3.1 Problem statement and notation

Fix a sparse Mixture-of-Experts (MoE) layer with n experts $\{f_1, \dots, f_n\}$ and a learned router g, and write $f(\cdot \mid x)$ for its output next-token distribution at input x. Given a target survivor count $k < n$, learning-free expert compression seeks a subset $S \subset \{1, \dots, n\}$ with $|S| = k$ and a router-redirect map $\pi : \{1, \dots, n\} \setminus S \to S$ such that the post-compression layer f_S minimizes the calibration-set KL divergence

$$\mathcal{L}(S, \pi) \;=\; \mathbb{E}_{x \sim \mathcal{D}}\, D_{\mathrm{KL}}\big(f(\cdot \mid x) \,\big\|\, f_S(\cdot \mid x)\big)\,, \tag{1}$$

under the constraint that f_S is obtained by offline computation on a small held-out calibration corpus $\mathcal{D}$ (App. B.11), with no fine-tuning, knowledge distillation, or LoRA adaptation.

We treat experts as vertices of an undirected complete graph $V = \{1, \dots, n\}$. For an unordered pair $\{i, j\} \in \binom{V}{2}$, the *pairwise merge barrier*

$$b_{ij} \;=\; \mathbb{E}_{x \sim \mathcal{D}}\, D_{\mathrm{KL}}\big(f(\cdot \mid x) \,\big\|\, f_{V \setminus \{i,j\} \cup \{i \oplus j\}}(\cdot \mid x)\big) \tag{2}$$

measures the distributional cost of replacing $\{i, j\}$ with their frequency-weighted merge $i \oplus j$, where the frequency weighting is the empirical token-routing rate of each expert on $\mathcal{D}$ and a guard falls back to the unweighted average when both rates vanish (App. B.10). For an unordered triple $\{i, j, k\} \in \binom{V}{3}$, the *triplet merge barrier* b_{ijk} is defined analogously, by replacing $\{i, j, k\}$ with their joint frequency-weighted merge $i \oplus j \oplus k$.

3.2 The mergeability complex and the simplicial Laplacian

The *mergeability complex* of an MoE layer is the abstract simplicial 2-complex $K = (V, E, T)$ with $E = \binom{V}{2}$ (the complete pairwise edge set) and $T \subseteq \binom{V}{3}$ a curated triangle set whose two-stage construction is included in App. B.9. Pairwise barriers form an edge-supported signal $b \in C_1(K)$ with coefficient b_{ij} on edge $\{i, j\}$, and triplet barriers form a triangle-supported signal $c \in C_2(K)$ with coefficient b_{ijk} on triangle $\{i, j, k\}$. We work with the complete edge set rather than a thresholded subgraph $\{(i, j) : b_{ij} \le \tau_e\}$, because thresholding distorts the Hodge spectrum: it changes the rank of ∂_1 on disconnection events and can detach otherwise harmonic cycles (App. B.8).

Fix the total ordering of V given by the checkpoint's expert indices and the induced lexicographic orientation on edges and triangles (App. B.3). Let $C_q(K) = \mathbb{R}^{F_q(K)}$, where $F_q(K)$ is the set of oriented q-simplices ($q = 0, 1, 2$); here q indexes the chain dimension, not the percentile thresholds

of Section 4. For oriented edges $[i, j]$ with $i < j$ and oriented triangles $[i, j, k]$ with $i < j < k$, the signed boundary operators are

$$\partial_1 : C_1(K) \to C_0(K), \qquad \partial_1 [i, j] = [j] - [i], \tag{3}$$

$$\partial_2 : C_2(K) \to C_1(K), \qquad \partial_2 [i, j, k] = [j, k] - [i, k] + [i, j], \tag{4}$$

satisfying the chain-complex identity $\partial_1 \circ \partial_2 = 0$ (Hatcher, 2002; Munkres, 1996). The *1-Hodge Laplacian* of K is

$$L_1 \;=\; \partial_1^\top \partial_1 \;+\; \partial_2 \, \partial_2^\top, \tag{5}$$

a symmetric positive semi-definite operator on $C_1(K)$ (Eckmann, 1944; Lim, 2020; Schaub et al., 2020; Horak & Jost, 2013). We use the unweighted L_1: barriers enter only as the signal $b \in C_1(K)$, never as weights inside the operator. The kernel $\ker(L_1)$ depends only on the combinatorial topology of K (App. B.3), so the unweighted L_1 keeps the harmonic projector independent of how barriers are normalized across layers and model families (Remark 3 treats weighted operators).

3.3 Hodge decomposition of the merge-barrier signal

Theorem 1 (Discrete Hodge decomposition; (Eckmann, 1944; Lim, 2020))**.** *Every edge-supported signal $b \in C_1(K)$ admits a unique orthogonal decomposition*

$$b \;=\; b_{\mathrm{grad}} \;+\; b_{\mathrm{curl}} \;+\; b_{\mathrm{harm}}, \tag{6}$$

with $b_{\mathrm{grad}} \in \mathrm{im}(\partial_1^\top)$, $b_{\mathrm{curl}} \in \mathrm{im}(\partial_2)$, and $b_{\mathrm{harm}} \in \ker(L_1)$. The three subspaces are pairwise orthogonal in $\mathbb{R}^{|E|}$, and $\dim \ker(L_1) = \beta_1(K)$, the first Betti number of K.

A self-contained linear-algebraic proof using the chain identity $\partial_1 \partial_2 = 0$ is in App. B.4; the explicit projection formulas $P_{\mathrm{grad}} = \partial_1^\top (\partial_1 \partial_1^\top)^+ \partial_1$, $P_{\mathrm{curl}} = \partial_2 (\partial_2^\top \partial_2)^+ \partial_2^\top$, and $P_{\mathrm{harm}} = I - P_{\mathrm{grad}} - P_{\mathrm{curl}}$, with $(\cdot)^+$ the Moore-Penrose pseudoinverse, are derived in App. B.5.

Proposition 1 (Harmonic energy as irreducible mergeability residual)**.** *Let $\mathcal{M}_K := \mathrm{im}(\partial_1^\top) \oplus \mathrm{im}(\partial_2) \subseteq C_1(K)$ denote the subspace of edge-barrier signals expressible as a vertex potential lifted to edges plus a triangle-boundary correction. For every $b \in C_1(K)$,*

$$\inf_{m \in \mathcal{M}_K} \|b - m\|^2 \;=\; \|b_{\mathrm{harm}}\|^2, \tag{7}$$

with unique minimizer $m^\star = b_{\mathrm{grad}} + b_{\mathrm{curl}}$.

In particular, the harmonic energy fraction $\rho_{\mathrm{harm}}(\ell)$ (Eq. 8) is the fraction of layer-ℓ merge-barrier energy that no vertex-potential or triangle-boundary explanation can capture (App. B.6). Proposition 1 is unconditional. For the redirect proxy $\widehat{\mathcal{L}}(a) = \sum_{i \notin S} b_{i,\pi(i)} = \langle w_a, b \rangle$ of a plan $a = (S, \pi)$, orthogonality gives $\widehat{\mathcal{L}}(a) - \langle w_a, b_{\mathrm{grad}} + b_{\mathrm{curl}} \rangle = \langle P_{\mathrm{harm}} w_a, b_{\mathrm{harm}} \rangle$: the harmonic-blind part misprices a plan by the signed harmonic exposure of its redirect edges (App. B.7). In a layerwise model with top-1 routing, the model's quadratic cost equals $\langle w_a, b^{(2)} \rangle$, with routing-determined weights and quadratic barriers $b^{(2)}$ (Prop. 4; hypotheses in App. B.7); on the appendix's realizable same-fiber pairs $b \pm h$, selectors that see only b_{lo}, the routing, and expert norms cannot separate the pair, and their worst-case regret is at least half the h-exposure gap between plans optimal at b (Prop. 5); §5.3–§5.4 and App. F.4 show that covering harmonic-critical edges tracks compression quality.

Under the fixed reference orientation, b_{grad}, b_{curl}, and b_{harm} are the orthogonal projections of b onto $\mathrm{im}\, \partial_1^\top$, $\mathrm{im}\, \partial_2$, and $\ker L_1$. They provide vertex-, triangle-boundary-, and residual harmonic components of the oriented barrier cochain. By analogy, pairwise rankings (Lasby et al., 2025; Lee et al., 2025) act on vertex-level scores, and hypergraph cuts (Zhou et al., 2006) and triplet penalties (Schroff et al., 2015; Wang et al., 2014) on sums over triangles; the residual harmonic component is the object HodgeCover adds. Joint-merge costs enter via triplet barriers in $T^\star$.

3.4 Per-layer diagnostic across three MoE families

Two complementary per-layer diagnostics test whether the harmonic component is non-trivial in production sparse MoEs. Let $b^{(\ell)} \in \mathbb{R}^{|E|}$ denote the edge-supported pairwise-KL merge-barrier signal at layer ℓ. The first is the *harmonic energy fraction*

$$\rho_{\mathrm{harm}}(\ell) \;=\; \frac{\left\| P_{\mathrm{harm}}\, b^{(\ell)} \right\|^2}{\left\| b^{(\ell)} \right\|^2} \qquad (b^{(\ell)} \neq 0), \tag{8}$$

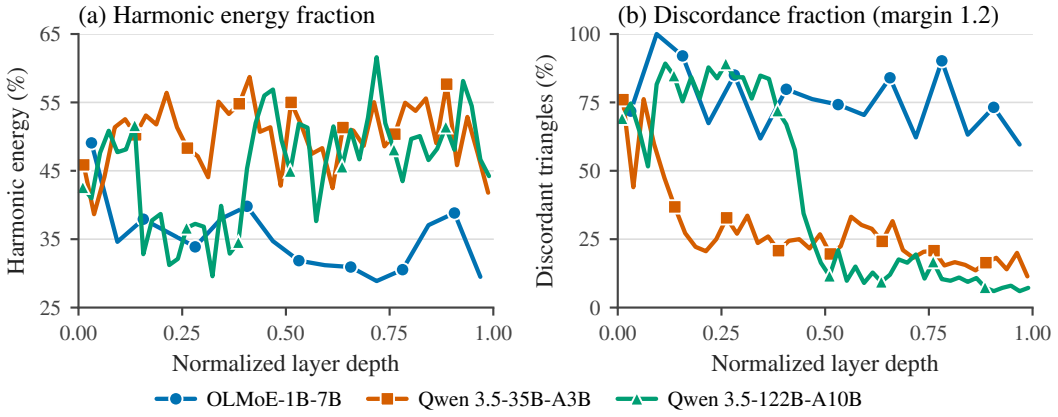

Figure 2: Harmonic energy fraction $\rho_{\text{harm}}(\ell)$ (Eq. 8) and discordance fraction $\delta(\ell)$ (Eq. 9) at every layer of three production MoE families. Both signals stay non-trivial throughout depth.

a global property of the simplicial Laplacian. The second is the combinatorial *discordance fraction*

$$\delta(\ell) \;=\; \frac{1}{|T|}\,\Big|\big\{\{i,j,k\}\in T : b_{ijk}^{(\ell)} > 1.2\cdot\max\big(b_{ij}^{(\ell)}, b_{ik}^{(\ell)}, b_{jk}^{(\ell)}\big)\big\}\Big|, \tag{9}$$

the fraction of sampled triangles whose joint merge barrier exceeds the worst pairwise barrier by a 20% margin. Both live in $[0,1]$ and are scale-invariant per layer; ρ_{harm} is a spectral property tied to $\ker(L_1)$, while δ is a combinatorial stress test that does not require diagonalizing L_1.

Figure 2 reports both diagnostics on every layer of OLMoE-1B-7B (Muennighoff et al., 2024), Qwen3.5-35B-A3B and Qwen3.5-122B-A10B (Qwen Team, 2026). The harmonic energy fraction stays non-trivial at every layer of every model, with the two Qwen variants in a slightly higher and tighter band than OLMoE; the discordance fraction is also non-trivial throughout, with OLMoE the most discordant model and the two Qwen models tapering in deeper layers. The two diagnostics agree in direction and disagree in shape, giving independent evidence of an irreducible higher-order obstruction; per-layer ranges, gradient/curl companion curves, and the Euler–Poincaré formula for the first Betti number $\beta_1(K)$ are given in App. B.12 and B.3.

Figure 2 thus quantifies the barrier energy that no lower-order model explains; Section 4 turns its largest harmonic edges into coverage constraints for survivor selection.

4 Method

We turn the diagnostic of Section 3 into a learning-free expert-selection algorithm, *HodgeCover* (§4.1), and a hybrid *HodgeCover+Wanda* that composes it with off-the-shelf weight pruning (§4.2).

4.1 HodgeCover

Let $k < n$ be the target survivor count from the layer's n experts. HodgeCover selects a survivor set $S^\star \subseteq V$ of size k in three steps.

Critical Simplices. From the Hodge decomposition $b = b_{\text{grad}} + b_{\text{curl}} + b_{\text{harm}}$ (Theorem 1) we extract the *harmonic-critical edge set* $E^\star \subseteq E$ as the top-$p\%$ edges ranked by $|b_{\text{harm},e}|$ (the pairwise-cycle signal), and the *triplet-critical triangle set* $T^\star \subseteq T$ as the top-$q_T\%$ triangles ranked by raw $|b_{ijk}|$, which already mark where joint merging amplifies pairwise cost (App. B.9), so no 2-harmonic projection is needed. The percentile thresholds p and q_T are reported in App. D.

Coverage Objective. Let $N_E(i) = \{e \in E^\star : i \in e\}$ and $N_T(i) = \{\sigma \in T^\star : i \in \sigma\}$ be the per-expert incidence sets, and write $C_E(S) = \bigcup_{i\in S} N_E(i)$, $C_T(S) = \bigcup_{i\in S} N_T(i)$ for what a candidate $S \subseteq V$ covers. HodgeCover defines the selection objective

$$\Phi(S) \;:=\; \underbrace{\sum_{i\in S}\text{sal}(i)}_{\text{saliency}} \;+\lambda_e\; \underbrace{\frac{|C_E(S)|}{|E^\star|}}_{\text{harmonic-edge coverage}} \;+\lambda_t\; \underbrace{\frac{|C_T(S)|}{|T^\star|}}_{\text{triplet-triangle coverage}}, \tag{10}$$

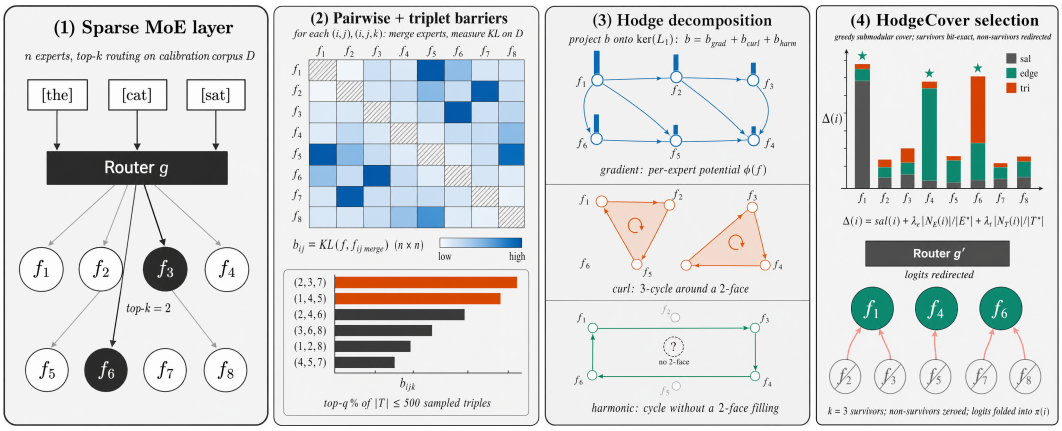

Figure 3: The four stages of HodgeCover: token-level routing, pairwise + triplet KL barriers on the calibration corpus, Hodge decomposition of the edge-supported barrier signal, and submodular survivor selection with a Hodge-weighted router redirect.

with $\lambda_e, \lambda_t \geq 0$ and saliencies $\mathrm{sal}(i) \in [0,1]$ given by the REAP output-saliency score (Lasby et al., 2025) normalized per layer (with the convention $|C_E(S)|/|E^\star| = 0$ when $E^\star = \emptyset$, and likewise for $T^\star$). The objective Φ is non-negative monotone submodular (Proposition 2); the standard greedy procedure adds, at each step, the expert maximizing the marginal score

$$\Delta(i \mid S) \;=\; \mathrm{sal}(i) \;+\; \lambda_e \frac{|N_E(i) \setminus C_E(S)|}{|E^\star|} \;+\; \lambda_t \frac{|N_T(i) \setminus C_T(S)|}{|T^\star|}, \tag{11}$$

and terminates after k additions. The denominators $|E^\star|, |T^\star|$ keep all three terms in $[0,1]$ across layers and models, so λ_e, λ_t are layer-agnostic. The harmonic-edge term targets the irreducible residual identified by Proposition 1; the triplet-triangle term protects high-cost joint merges that enter through the curated triangle set $T^\star$.

Proposition 2 (Topological coverage of HodgeCover)**.** *The HodgeCover selection objective Φ in Eq. 10 is a non-negative monotone submodular set function in $S \subseteq V$ subject to the cardinality constraint $|S| = k$, and the greedy survivor set $S^\star$ produced by iterating Eq. 11 from the empty set (*$\mathrm{SE} = \emptyset$ *in Algorithm 1, the setting used for every reported run) satisfies*

$$\Phi(S^\star) \;\geq\; \big(1 - 1/e\big) \max_{S \subseteq V,\, |S| = k} \Phi(S). \tag{12}$$

Φ is a surrogate; App. C.2 proves this and shows that, in the quadratic model, an extended redirect-or-zero action class has a monotone-submodular facility-location output-error reduction, and gives explicit comparison conditions under which the guarantee for Φ would transfer to that reduction.

Remark 1 (Non-modularity of coverage)**.** *Coverage is non-modular because selected experts can cover the same critical simplex, so $|C_E(S)|$ is not $\sum_{i \in S} h(i)$ for any per-expert score $h : V \to \mathbb{R}$. Rankings such as REAP (Lasby et al., 2025), REAM centroid selection (Jha et al., 2026), and MC-SMoE (Li et al., 2023) apply a top-k rule to a fixed vertex score, and a static incident-mass score need not reproduce the greedy marginal gains of Eq. 11; App. C.2 gives an example.*

Router Redirect for non-Survivors. Let $N^\star = V \setminus S^\star$ be the dropped experts. We use the REAP redirect surgery (Lasby et al., 2025): expert weights of each $i \in N^\star$ are erased and its router logit is folded into a single survivor $\pi(i) \in S^\star$, so the gating distribution renormalizes onto $S^\star$ at no extra forward pass. Where REAP picks the cosine-nearest survivor in router-key space, we pick the nearest survivor under a *Hodge-weighted barrier* that up-weights harmonic edges:

$$\pi(i) \;=\; \arg\min_{j \in S^\star} \; b_{ij} \cdot \Big(1 + \alpha \, \tfrac{|b_{\mathrm{harm},\{i,j\}}|}{\|b\|}\Big), \qquad i \in N^\star, \tag{13}$$

where the strength $\alpha > 0$ controls how much edges that carry their own harmonic mass are penalized, thereby biasing redirects away from such edges, subject to the barrier tradeoff. We adopt the convention that the harmonic factor is 1 when $\|b\| = 0$. Survivors are never perturbed: their gate

vectors and expert weights are kept identical, so the post-compression layer is the survivor sub-MoE with a renormalized router. These three steps run independently per MoE layer; the cross-layer budget allocator that turns a global compression rate into the per-layer counts $\{k_\ell\}_\ell$ is given in App. C.

Costs. The HodgeCover pipeline is offline and learning-free. Plan-time is dominated by the $O\big((n^2 + |T|)\,|\mathcal{D}|\,d^2\big)$ barrier sweep (d = expert hidden dim, $|\mathcal{D}|$ = calibration tokens); the complex is computed once per model and cached across compression rates. On OLMoE and Qwen 3.5-35B, the measured first-run plan-time of HodgeCover and HodgeCover+Wanda lies within the range spanned by the learning-free baselines (Table 14), and post-compression throughput of HodgeCover+Wanda at 66% exceeds REAP+Wanda's on OLMoE and Qwen 3.5-122B and is within 6% on every backbone (Table 15). Algorithm pseudocode, hyperparameter defaults, the cross-layer allocator, and a baseline-axis comparison grid (Table 3) versus the other learning-free MoE compressors benchmarked in §5 are in App. C and D.

4.2 HodgeCover+Wanda

HodgeCover targets the expert-count axis; we obtain a hybrid compressor, *HodgeCover+Wanda*, by composing it with an off-the-shelf one-shot weight pruner. Stage 1 runs HodgeCover at a fixed expert-count drop rate r_1, yielding the survivor set $S^\star$ and the redirected router. Stage 2 applies unstructured Wanda (Sun et al., 2023) on the survivor weights at the residual sparsity needed to reach the total target rate, reusing the same calibration corpus $\mathcal{D}$ at no additional forward-pass. Stage 1 is the contribution; Stage 2 is plug-and-play (any calibration-aware unstructured weight pruner could replace Wanda). Pseudocode, the protocol that selects r_2 given r_1, and the per-cell values of r_1, r_2 are in App. C.1 and C.4.

5 Experiments

5.1 Setup

We benchmark three open-weight sparse MoE backbones spanning two expert-count regimes and two scales: OLMoE-1B-7B (Muennighoff et al., 2024) (16 MoE layers, 64 experts), Qwen 3.5-35B-A3B and Qwen 3.5-122B-A10B (Qwen Team, 2026) (40 / 48 MoE layers, 256 experts). Each backbone is compressed at 33% and 66% of expert count. REAP (Lasby et al., 2025), REAM (Jha et al., 2026), and STUN+Wanda (Lee et al., 2025) run their published implementations on a shared $2{,}048$-token calibration footing; MC-SMoE (Li et al., 2023) runs in its learning-free merging configuration; REAP+Wanda and REAM+Wanda are matched controls we construct (App. C.7). HodgeCover (ours) uses the same uniform per-layer convention as REAP and REAM (App. F.1). Calibration uses $2{,}048$ C4-train tokens (Raffel et al., 2020); evaluation reports WikiText-103 perplexity (Merity et al., 2016), C4 perplexity, and nine downstream tasks via the LM Evaluation Harness (Gao et al., 2021): ARC-c and ARC-e (Clark et al., 2018), BoolQ (Clark et al., 2019), HellaSwag (Zellers et al., 2019), MMLU 5-shot (Hendrycks et al., 2020), PIQA (Bisk et al., 2020), TruthfulQA-MC2 (Lin et al., 2022), WinoGrande (Sakaguchi et al., 2021), and GSM8K 8-shot (Cobbe et al., 2021); the *downstream-task average* (DS-Avg) is their unweighted mean. Beyond the four published baselines we also report four ablations of HodgeCover that share its calibration corpus and per-layer expert budget but replace the topological objective with a no-triangle, soft-triplet-penalty, hard-triplet-veto, or topology-blind alternative (Zhou et al., 2006); precise constructions, hardware, harness configuration, and per-method allocator descriptions are in App. F.1 and App. E.

5.2 Main results

HodgeCover+Wanda has the best C4 perplexity on every (model, rate) cell in Table 1 and the best WikiText perplexity on every 66% cell, with WikiText margins of 1.4–3.8 points over STUN+Wanda at 66% on the three backbones; at 33%, STUN+Wanda is ahead on WikiText for the two smaller backbones (OLMoE and Qwen 3.5-35B-A3B) at the cost of 1.7–2.2 C4 perplexity points. HodgeCover+Wanda leads STUN+Wanda in DS-Avg on all four Qwen (model, rate) cells; at 66% on Qwen 3.5-35B it gains $+12.6$ pp DS-Avg over STUN+Wanda (74.6% vs. 62.0%), driven primarily by GSM8K (85.5% vs. 22.1%); the same pattern holds at 66% on Qwen 3.5-122B-A10B. On the expert-count axis, HodgeCover is within ± 0.3 pp DS-Avg of REAP on all four Qwen cells, while leading REAM and MC-SMoE by 6–16 pp at 66%. Paired per-example tests on matched re-runs reject the null hypothesis at all five tested 66% contrasts with HodgeCover ahead, and calibration-seed replicates on OLMoE keep its margin over REAP positive at every seed (App. F.3). At 66% on OL-

Table 1: Per-task downstream accuracy (%) plus WikiText-103 / C4 perplexity, with the rightmost column reporting the unweighted nine-task DS-Avg. **Bold** = best in column within the same axis; underline = second-best in the expert-count axis.

Model	Rate	Method	Wiki	C4	ARC-c	ARC-e	BoolQ	HellaS	MMLU	PIQA	TQA	WinoG	GSM8K	DS-Avg
OLMoE 1B-7B	–	*Uncompressed*	*8.17*	*13.64*	*48.9*	*76.2*	*75.0*	*77.0*	*52.8*	*80.7*	*35.6*	*68.7*	*9.8*	*58.3*
	33%	*Expert-count reduction methods*												
		REAP	**19.28**	**19.62**	28.7	42.5	**65.3**	63.8	27.4	66.8	35.5	64.8	1.7	44.1
		REAM	33.00	26.64	27.1	39.3	63.3	53.0	**27.5**	63.4	41.5	60.9	1.3	41.9
		MC-SMoE	39.56	39.87	26.3	41.8	62.3	43.3	27.2	65.1	39.2	50.3	1.2	39.6
		HodgeCover (ours)	23.50	22.91	**32.8**	**50.3**	62.8	**67.0**	25.6	**73.0**	**41.7**	**64.9**	**2.2**	**46.7**
		Hybrid methods												
		STUN+Wanda	**11.85**	18.43	**41.5**	67.2	66.4	69.2	34.6	77.3	33.3	64.8	**2.6**	50.8
		HodgeCover+Wanda (ours)	13.92	**16.69**	40.9	**67.3**	**73.0**	**73.6**	**38.7**	**78.1**	**35.1**	**67.8**	2.4	**53.0**
	66%	*Expert-count reduction methods*												
		REAP	**864.7**	**180.8**	**24.5**	**30.9**	**48.0**	**32.3**	25.1	**54.2**	**50.7**	**53.2**	0.0	**35.4**
		REAM	4,131.2	1,026.3	24.3	30.4	42.8	28.8	25.3	53.2	50.2	50.9	0.0	34.0
		MC-SMoE	1,443.7	835.2	23.6	30.8	43.3	27.9	23.5	54.1	50.0	50.8	0.0	33.8
		HodgeCover (ours)	1,134.9	535.5	24.2	29.5	38.6	30.0	**25.4**	52.7	47.7	51.2	**0.1**	33.3
		Hybrid methods												
		STUN+Wanda	22.11	31.01	32.9	**58.0**	62.5	55.3	26.9	71.7	**36.0**	61.1	1.8	45.1
		HodgeCover+Wanda (ours)	**18.32**	**21.18**	**33.7**	56.6	**72.6**	**64.4**	**35.0**	**72.6**	35.2	**64.5**	**2.5**	**48.6**
Qwen 3.5 35B-A3B	–	*Uncompressed*	*7.25*	*13.19*	*62.1*	*79.5*	*87.9*	*82.5*	*84.9*	*83.0*	*53.5*	*74.7*	*86.9*	*77.2*
	33%	*Expert-count reduction methods*												
		REAP	**9.65**	**13.71**	**57.1**	**76.0**	90.0	81.5	79.8	82.0	53.2	75.1	**87.6**	75.8
		REAM	9.91	14.14	54.4	73.1	90.5	80.6	78.8	81.2	53.8	74.0	86.8	74.8
		MC-SMoE	10.90	14.97	45.1	64.6	89.2	78.3	62.5	81.0	51.3	76.1	83.9	70.2
		HodgeCover (ours)	9.97	13.75	56.3	73.9	**91.3**	**82.5**	**80.7**	**82.9**	**54.7**	**76.3**	85.1	**75.9**
		Hybrid methods												
		STUN+Wanda	**8.86**	15.61	50.9	74.4	**89.2**	72.1	76.7	76.1	51.0	70.6	33.2	66.0
		HodgeCover+Wanda (ours)	9.18	**13.42**	**55.9**	**76.0**	89.0	**81.8**	**82.6**	**82.3**	**53.7**	**74.2**	**89.6**	**76.1**
	66%	*Expert-count reduction methods*												
		REAP	**14.97**	19.29	46.0	**69.8**	**88.2**	71.4	**58.7**	77.3	48.2	72.5	65.3	66.4
		REAM	16.80	22.32	42.7	64.4	87.2	67.9	51.7	74.1	46.9	70.9	39.0	60.5
		MC-SMoE	20.64	24.25	29.0	47.5	85.1	54.9	32.7	69.4	47.2	67.6	34.6	52.0
		HodgeCover (ours)	15.13	**18.86**	**46.4**	67.5	88.1	**72.5**	54.5	**78.1**	**51.7**	**73.7**	**67.5**	**66.7**
		Hybrid methods												
		STUN+Wanda	11.77	20.05	49.6	76.6	89.5	63.8	69.2	72.9	47.1	67.5	22.1	62.0
		HodgeCover+Wanda (ours)	**10.25**	**15.22**	**56.5**	**78.5**	**90.8**	**76.8**	**78.0**	**80.8**	**50.3**	**74.1**	**85.5**	**74.6**
Qwen 3.5 122B-A10B	–	*Uncompressed*	*4.50*	*12.33*	*63.6*	*80.5*	*86.3*	*85.9*	*88.1*	*83.4*	*51.8*	*74.6*	*88.3*	*78.0*
	33%	*Expert-count reduction methods*												
		REAP	7.21	**12.78**	**63.3**	81.6	71.8	84.8	**84.7**	**83.3**	53.2	77.1	81.7	**75.7**
		REAM	7.52	12.98	61.9	**81.7**	74.6	83.5	83.5	83.1	52.9	75.6	82.6	75.5
		MC-SMoE	9.44	13.96	49.2	72.4	**80.5**	81.8	71.6	82.1	**54.0**	75.6	82.4	72.2
		HodgeCover (ours)	**7.21**	12.81	61.7	80.1	74.4	**84.9**	83.9	83.1	51.6	**77.3**	**84.5**	75.7
		Hybrid methods												
		STUN+Wanda	6.35	14.48	**62.5**	**84.6**	**88.8**	78.8	83.1	79.8	**52.6**	74.6	74.8	75.5
		HodgeCover+Wanda (ours)	**5.99**	**12.49**	62.2	79.9	84.2	**85.6**	**85.5**	**83.4**	51.2	**77.5**	**90.3**	**77.8**
	66%	*Expert-count reduction methods*												
		REAP	**12.24**	17.28	50.2	**71.3**	77.9	75.1	**70.0**	79.8	50.5	75.0	**80.2**	70.0
		REAM	13.94	19.67	46.8	67.6	83.7	70.5	65.4	75.5	48.8	73.6	30.4	62.5
		MC-SMoE	18.21	22.82	33.6	50.5	**86.0**	62.9	34.6	72.4	44.0	72.5	34.0	54.5
		HodgeCover (ours)	12.46	**17.21**	**52.2**	71.1	75.3	**77.0**	69.0	**80.1**	**51.5**	**76.4**	78.1	**70.1**
		Hybrid methods												
		STUN+Wanda	8.77	17.50	59.6	**84.8**	**82.1**	71.0	78.3	76.9	49.6	73.0	62.3	70.8
		HodgeCover+Wanda (ours)	**7.42**	**13.86**	**61.2**	82.3	79.3	**80.1**	**83.2**	**81.9**	**51.1**	**76.4**	**87.9**	**75.9**

MoE, HodgeCover+Wanda reaches 21.2 C4 perplexity and 48.6% DS-Avg, $+3.5$ pp over STUN+W, in a regime where every expert-count method crosses C4 perplexity 100 and the uniform per-layer budget falls below the critical-simplex hitting-set minimum (App. F.5). The matched-control hybrids REAP+Wanda and REAM+Wanda trail HodgeCover+W at 66% by 0.3–0.8 pp DS-Avg on the two Qwen scales and by 5–8 pp DS-Avg on OLMoE (App. F.2).

5.3 Mechanism: how baselines deviate from HodgeCover

For each Hodge component and for the triplet barriers we measure the fraction of uncompressed ℓ^1 mass retained by the survivor set (Eq. 35, App. F.6) and report each baseline's deviation from HodgeCover. Figure 4 resolves the baselines into two trade-off classes. Greedy-Barrier and MC-SMoE retain 0.05–0.07 more harmonic and gradient mass at the cost of 0.10–0.25 less curl and triplet-barrier mass; Random sweeps the opposite trade, gaining curl and triplet-barrier mass while losing harmonic and gradient. The HodgeCover coverage objective Φ does not pick a side: by greedy-covering the top-$p\%$ harmonic-critical edges and top-$q_T\%$ triplet-critical triangles jointly, it stays close to the backbone's pre-compression decomposition along all four components, consistent with its 6–16 pp DS-Avg lead over Greedy-Barrier and MC-SMoE on the two Qwen scales. REAP and REAM share identical pure-axis survivor selection; their deviation from HodgeCover is weak

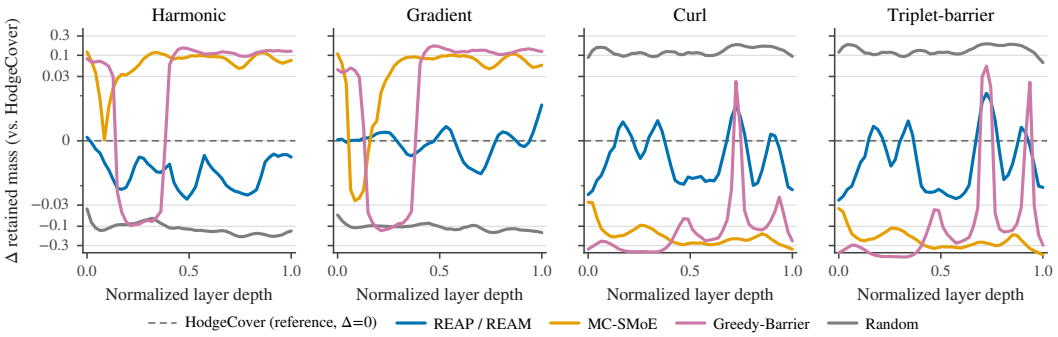

Figure 4: Per-layer deviation from HodgeCover on Qwen 3.5-122B-A10B at 66% across the four Hodge components. Greedy-Barrier and MC-SMoE pay for higher harmonic and gradient retention with sharply lower curl and triplet-barrier mass; Random sweeps in the opposite direction; REAP / REAM (whose pure-axis survivor selection is identical) deviate weakly on every component.

across all four components, consistent with their narrow DS-Avg gap to HodgeCover in Table 1. Matched plots on the other two backbones and macro numerics are in App. F.6.

5.4 Ablation Study

We compare HodgeCover against the four ablations of Section 5.1 on Qwen 3.5-35B-A3B (Table 2). Triplet-Penalty consumes the same triplet table through a sum-of-barriers penalty with merge-based survivor treatment (App. C.7); routing them through the Hodge coverage objective instead gains $+5.7$ pp DS-Avg at 66% at near-identical C4 perplexity (18.86 vs. 18.77). Removing triangles costs -11.5 pp DS-Avg, and replacing the soft objective with a binary triangle veto collapses by -30.5 pp, so both the triangle term and its soft formulation are necessary (per-task results: App. F.7; systems: App. F.9). Varying one hyperparameter at a time over $|T|$, p, q_T, (λ_e, λ_t), triangle seeds, and an alternative triangle sampler moves DS-Avg by 0.22 to 1.52 pp at non-collapsed cells, with survivor-set Jaccard 0.999; calibration size moves DS-Avg by about 3 pp (App. D.2). An additive per-expert selector built on the same harmonic-critical and triplet-critical quantities recovers most of HodgeCover's survivors; HodgeCover is ahead of it in five of seven cells and within 0.16 pp in the other two (App. F.4). The gain therefore comes from the Hodge-derived quantities, and greedy coverage optimizes them with the $(1 - 1/e)$ guarantee.

Table 2: Ablations on Qwen 3.5-35B-A3B. **Bold** = best in column.

Variant	Wiki	C4	DS-Avg
33% *rate*			
HodgeCover (ours)	**9.97**	**13.75**	**75.9**
Hodge No-Triangle	12.48	17.10	67.7
Triplet-Hypergraph	14.29	16.76	64.5
Triplet-Penalty	10.22	13.82	72.3
Greedy-Barrier	10.22	13.79	72.7
66% *rate*			
HodgeCover (ours)	**15.13**	18.86	**66.7**
Hodge No-Triangle	21.34	28.34	55.2
Triplet-Hypergraph	98.08	154.9	36.2
Triplet-Penalty	16.14	18.77	60.9
Greedy-Barrier	15.30	**18.13**	60.7

6 Conclusion

Pairwise merge scores cannot represent a higher-order property of MoE mergeability: three experts can be pairwise compatible yet collectively form an irreducible cycle. The Hodge decomposition of the simplicial mergeability complex isolates this cycle as a harmonic component, and covering the harmonic-critical edges and triplet-critical triangles with a learning-free objective gives the best 66% hybrid results across three MoE backbones.

Limitations and Broader Impact. HodgeCover stays inside the learning-free family: recovering the last few perplexity points typically requires fine-tuning or distillation, which is orthogonal to our pipeline. The barrier sweep is quadratic in the expert count and has been run up to 256 experts per layer; on OLMoE and Qwen 3.5-35B its first-run cost lies within the range of the listed baselines (Table 14). The construction consumes only routing statistics and expert outputs, but the main evaluation is language-only on base checkpoints; App. F.8 adds one preliminary vision-language backbone at one rate. As with other compression methods, lowering the inference cost of pretrained MoE models without retraining both widens access and accelerates deployment of insufficiently aligned checkpoints; the practitioner remains responsible for auditing any checkpoint they deploy.

AI USE STATEMENT

We used generative AI tools to aid and polish the writing of this paper, namely language editing, phrasing, and tightening of the manuscript text. The research ideas, theory, experiments, and analysis are the authors' own. The authors reviewed and edited all AI-assisted text, and take full responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

This work compresses publicly released pretrained sparse MoE checkpoints using public calibration and evaluation datasets. It involves no human subjects, no personal data, and no new data collection. HodgeCover lowers the inference cost of existing models without retraining; as with other post-training compression methods, this both widens access and can accelerate deployment of insufficiently aligned checkpoints, so the practitioner remains responsible for auditing any checkpoint they compress and deploy. We are not aware of conflicts of interest or sponsorship concerns beyond standard academic funding, and the work adheres to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

The anonymized implementation of HodgeCover and HodgeCover+Wanda, with setup and reproduction instructions, is linked in App. A. Models, compression rates, the calibration corpus, the evaluation harness configuration, and hardware are specified in App. E; hyperparameter defaults and how they were set are in App. D.1; implementation notes for every baseline, including each deviation from its source paper, are in App. C.7. Complete proofs of the stated results are in App. B and App. C.2.

REFERENCES

Sameer Agarwal, Kristin Branson, and Serge Belongie. Higher order learning with graphs. In *Proceedings of the 23rd international conference on Machine learning*, pp. 17–24, 2006.

Song Bai, Feihu Zhang, and Philip HS Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110:107637, 2021.

Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Cristian Bodnar, Fabrizio Frasca, Yuguang Wang, Nina Otter, Guido F Montufar, Pietro Lio, and Michael Bronstein. Weisfeiler and lehman go topological: Message passing simplicial networks. In *International conference on machine learning*, pp. 1026–1037. PMLR, 2021.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

I Chen, Hsu-Shen Liu, Wei-Fang Sun, Chen-Hao Chao, Yen-Chang Hsu, Chun-Yi Lee, et al. Retraining-free merging of sparse moe via hierarchical clustering. *arXiv preprint arXiv:2410.08589*, 2024.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, pp. 2924–2936, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. *arXiv preprint arXiv:2409.17146*, 2024.

Stefania Ebli, Michaël Defferrard, and Gard Spreemann. Simplicial neural networks. *arXiv preprint arXiv:2010.03633*, 2020.

Beno Eckmann. Harmonische funktionen und randwertaufgaben in einem komplex. *Commentarii Mathematici Helvetici*, 17(1):240–255, 1944.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3558–3565, 2019.

Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International conference on machine learning*, pp. 10323–10337. PMLR, 2023.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Joel Friedman. Computing betti numbers via combinatorial laplacians. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of Computing*, pp. 386–391, 1996.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. A framework for few-shot language model evaluation. *Zenodo*, 2021.

Timothy E Goldberg. *Combinatorial Laplacians of simplicial complexes*. PhD thesis, Bard College, 2002.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, 2017.

Vincent P Grande and Michael T Schaub. Disentangling the spectral properties of the hodge laplacian: not all small eigenvalues are equal. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9896–9900. IEEE, 2024.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

Allen Hatcher. *Algebraic Topology*. Cambridge University Press, Cambridge, UK, 2002. ISBN 978-0-521-79540-1.

Shwai He, Daize Dong, Liang Ding, and Ang Li. Towards efficient mixture of experts: A holistic study of compression techniques. *arXiv preprint arXiv:2406.02500*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Danijela Horak and Jürgen Jost. Spectra of combinatorial laplace operators on simplicial complexes. *Advances in Mathematics*, 244:303–336, 2013.

Max Horn, Edward De Brouwer, Michael Moor, Yves Moreau, Bastian Rieck, and Karsten Borgwardt. Topological graph neural networks. *arXiv preprint arXiv:2102.07835*, 2021.

Saurav Jha, Maryam Hashemzadeh, Ali Saheb Pasand, Ali Parviz, Min-Joong Lee, and Boris Knyazev. Ream: Merging improves pruning of experts in llms. *arXiv preprint arXiv:2604.04356*, 2026.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training mixture-of-experts from dense checkpoints. *arXiv preprint arXiv:2212.05055*, 2022.

Mike Lasby, Ivan Lazarevich, Nish Sinnadurai, Sean Lie, Yani Ioannou, and Vithursan Thangarasa. Reap the experts: Why pruning prevails for one-shot moe compression. *arXiv preprint arXiv:2510.13999*, 2025.

Jaeseong Lee, Seung-won Hwang, Aurick Qiao, Daniel F Campos, Zhewei Yao, and Yuxiong He. Stun: Structured-then-unstructured pruning for scalable moe pruning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13660–13676, 2025.

James R Lee, Shayan Oveis Gharan, and Luca Trevisan. Multiway spectral partitioning and higher-order cheeger inequalities. *Journal of the ACM (JACM)*, 61(6):1–30, 2014.

Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A Smith, and Luke Zettlemoyer. Branch-train-merge: Embarrassingly parallel training of expert language models. *arXiv preprint arXiv:2208.03306*, 2022.

Pingzhi Li, Zhenyu Zhang, Prateek Yadav, Yi-Lin Sung, Yu Cheng, Mohit Bansal, and Tianlong Chen. Merge, then compress: Demystify efficient smoe with hints from its routing policy. *arXiv preprint arXiv:2310.01334*, 2023.

Lek-Heng Lim. Hodge laplacians on graphs. *Siam Review*, 62(3):685–715, 2020.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6: 87–100, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

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. Springer, 2014.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Enshu Liu, Junyi Zhu, Zinan Lin, Xuefei Ning, Matthew B Blaschko, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. Efficient expert pruning for sparse mixture-of-experts language models: Enhancing performance and reducing inference costs. *arXiv preprint arXiv:2407.00945*, 2024b.

Zongfang Liu, Shengkun Tang, Yifan Shen, Huan Wang, and Xin Yuan. Aimer: Calibration-free task-agnostic moe pruning. *arXiv preprint arXiv:2603.18492*, 2026a.

Zongfang Liu, Shengkun Tang, Boyang Sun, Zhiqiang Shen, and Xin Yuan. Evoesap: Non-uniform expert pruning for sparse moe. *arXiv preprint arXiv:2603.06003*, 2026b.

Anand Louis. Hypergraph markov operators, eigenvalues and approximation algorithms. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 713–722, 2015.

Xudong Lu, Qi Liu, Yuhui Xu, Aojun Zhou, Siyuan Huang, Bo Zhang, Junchi Yan, and Hongsheng Li. Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6159–6172, 2024.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720, 2023.

Xin Men, Mingyu Xu, Qingyu Zhang, Qianhao Yuan, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Shortgpt: Layers in large language models are more redundant than you expect. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20192–20204, 2025.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, et al. Olmoe: Open mixture-of-experts language models. *arXiv preprint arXiv:2409.02060*, 2024.

J.R. Munkres. *Elements Of Algebraic Topology*. Avalon Publishing, 1996. ISBN 9780201627282.

Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim. Topology of deep neural networks. *Journal of Machine Learning Research*, 21(184):1–40, 2020.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

Qwen Team. Qwen3.5: Towards native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`, February 2026. Alibaba Cloud / Qwen team blog post.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Bastian Rieck, Matteo Togninalli, Christian Bock, Michael Moor, Max Horn, Thomas Gumbsch, and Karsten Borgwardt. Neural persistence: A complexity measure for deep neural networks using algebraic topology. *arXiv preprint arXiv:1812.09764*, 2018.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Victor Sanh, Thomas Wolf, and Alexander Rush. Movement pruning: Adaptive sparsity by fine-tuning. *Advances in neural information processing systems*, 33:20378–20389, 2020.

Michael T Schaub, Austin R Benson, Paul Horn, Gabor Lippner, and Ali Jadbabaie. Random walks on simplicial complexes and the normalized hodge 1-laplacian. *SIAM Review*, 62(2):353–391, 2020.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. *arXiv preprint arXiv:2308.13137*, 2023.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Daniel A Spielman and Shang-Hua Teng. Spectral sparsification of graphs. *SIAM Journal on Computing*, 40(4):981–1025, 2011.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*, 2023.

Taiji Suzuki, Hiroshi Abe, Tomoya Murata, Shingo Horiuchi, Kotaro Ito, Tokuma Wachi, So Hirai, Masatoshi Yukishima, and Tomoaki Nishimura. Spectral pruning: Compressing deep neural networks via spectral analysis and its generalization error. *arXiv preprint arXiv:1808.08558*, 2018.

Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1386–1393, 2014.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, pp. 38087–38099. PMLR, 2023.

Yanyue Xie, Zhi Zhang, Ding Zhou, Cong Xie, Ziang Song, Xin Liu, Yanzhi Wang, Xue Lin, and An Xu. Moe-pruner: Pruning mixture-of-experts large language model using the hints from its router. *arXiv preprint arXiv:2410.12013*, 2024.

Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Yuanlin Duan, Wenqi Jia, Miao Yin, Yu Cheng, and Bo Yuan. Moe-i2: Compressing mixture of experts models through inter-expert pruning and intra-expert low-rank decomposition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10456–10466, 2024.

Lu Yin, You Wu, Zhenyu Zhang, Cheng-Yu Hsieh, Yaqing Wang, Yiling Jia, Gen Li, Ajay Jaiswal, Mykola Pechenizkiy, Yi Liang, et al. Outlier weighed layerwise sparsity (owl): A missing secret sauce for pruning llms to high sparsity. *arXiv preprint arXiv:2310.05175*, 2023.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Dongzhe Zheng, Tao Zhong, and Christine Allen-Blanchette. Topology-preserving neural operator learning via hodge decomposition. In *International conference on machine learning*. PMLR, 2026.

Tao Zhong, Zhixiang Chi, Li Gu, Yang Wang, Yuanhao Yu, and Jin Tang. Meta-dmoe: Adapting to domain shift by meta-distillation from mixture-of-experts. *Advances in Neural Information Processing Systems*, 35:22243–22257, 2022.

Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. *Advances in neural information processing systems*, 19, 2006.

A Code Availability

To facilitate the review process, we provide the full implementation of our method in an anonymized repository: `https://anonymous.4open.science/r/HodgeCover-18979`. Instructions for setting up and reproducing the proposed method are provided in the repository `README`.

B Hodge Decomposition Primer and Implementation Details

This appendix carries the full background and proofs deferred from Section 3. App. B.1 reviews the abstract simplicial complex; App. B.2 defines the boundary operators and proves the chain identity; App. B.3 constructs the combinatorial 1-Hodge Laplacian, characterizes its kernel, and gives the Euler–Poincaré formula for $\beta_1(K)$; App. B.4 states and proves the discrete Hodge decomposition; App. B.5 derives explicit projection formulas via the Moore-Penrose pseudoinverse; App. B.8 documents the effect of edge thresholding on the spectrum and motivates working with the complete edge set; App. B.9 analyzes the sensitivity of b_{harm} to the curated triangle set T; App. B.10 defines the frequency-weighted merge operation and its all-zero-frequency guard; App. B.11 fixes the calibration corpus; App. B.12 reports the per-layer ranges, gradient/curl companion curves, and the representative discordance margin used in Figure 2.

We follow the textbook treatments of Hatcher (2002); Munkres (1996) for classical homology, and Eckmann (1944); Lim (2020); Schaub et al. (2020); Horak & Jost (2013); Goldberg (2002); Friedman (1996); Zheng et al. (2026) for the discrete Hodge framework.

B.1 Abstract simplicial complexes

An *abstract simplicial complex* on a finite vertex set $V = \{1, \dots, n\}$ is a family $K \subseteq 2^V$ of non-empty subsets that is closed under taking non-empty subsets: $\sigma \in K$ and $\emptyset \neq \tau \subseteq \sigma$ implies $\tau \in K$. Elements $\sigma \in K$ with $|\sigma| = q + 1$ are *q-simplices*; the set of q-simplices is denoted $F_q(K)$. We work exclusively with $q \in \{0, 1, 2\}$: vertices, edges, triangles. Throughout this paper $K = (V, E, T)$ with $V = F_0(K) = \{1, \dots, n\}$, $E = F_1(K)$ the complete pairwise edge set $\binom{V}{2}$, and $T = F_2(K) \subseteq \binom{V}{3}$ a curated triangle set whose construction is given in App. B.9.

To do linear algebra on K we orient every simplex. Fix the total ordering $<$ on V given by the checkpoint's expert indices. For each simplex $\{v_0, \dots, v_q\} \in F_q(K)$ with $v_0 < v_1 < \cdots < v_q$, we choose the canonical lexicographically-ordered representative $[v_0, \dots, v_q]$ as its orientation; an odd permutation of these indices represents the opposite orientation. Equivalently, an *oriented q-simplex* is the equivalence class of an ordered $(q+1)$-tuple of distinct vertices modulo even permutations.

The space of *q-chains* is the real vector space

$$C_q(K) \;=\; \mathbb{R}^{F_q(K)},$$

with the standard basis indexed by oriented q-simplices. For $q = 0, 1, 2$ the dimensions are $\dim C_0 = |V| = n$, $\dim C_1 = |E|$, and $\dim C_2 = |T|$. We endow each $C_q(K)$ with the Euclidean inner product in the lexicographic basis. This is the unweighted choice; weighted variants are discussed in Remark 3.

B.2 Boundary operators and the chain identity

The boundary of an oriented simplex is the formal alternating sum of its oriented faces. Concretely,

$$\begin{aligned} &\partial_1 \colon C_1(K) \to C_0(K), && \partial_1\,[i,j] = [j] - [i], && i < j, \\ &\partial_2 \colon C_2(K) \to C_1(K), && \partial_2\,[i,j,k] = [j,k] - [i,k] + [i,j], && i < j < k. \end{aligned}$$

Extending linearly, ∂_1 is represented in the canonical bases by the signed vertex-edge incidence matrix $B_1 \in \{-1, 0, +1\}^{|V| \times |E|}$ of the underlying graph; ∂_2 is represented by the signed edge-triangle incidence matrix $B_2 \in \{-1, 0, +1\}^{|E| \times |T|}$.

Lemma 1 (Chain identity)**.** *$\partial_1 \circ \partial_2 = 0$ as a linear map $C_2(K) \to C_0(K)$. Equivalently,* $\mathrm{im}(\partial_2) \subseteq \ker(\partial_1)$.

Proof. By linearity it suffices to check the identity on a single basis element. Fix an oriented triangle $[i, j, k]$ with $i < j < k$. Then

$$\begin{aligned} \partial_1\,\partial_2\,[i,j,k] \;&=\; \partial_1\,\big([j,k] - [i,k] + [i,j]\big) \\ &=\; \big([k] - [j]\big) \;-\; \big([k] - [i]\big) \;+\; \big([j] - [i]\big) \\ &=\; [k] - [j] - [k] + [i] + [j] - [i] \\ &=\; 0, \end{aligned}$$

where every term cancels with its mate of opposite sign. □

The chain identity is the discrete counterpart of $d \circ d = 0$ for the exterior derivative, and it is the single algebraic fact that makes the Hodge decomposition possible.

Remark 2 (Why this is a 2-complex and not a hypergraph)**.** *A hypergraph treats each triangle $\{i, j, k\}$ as a single hyperedge with no internal structure. The simplicial complex K retains the boundary relation between an oriented triangle and its three oriented edges, which is exactly the structure that distinguishes "three pairwise compatibilities" from "a coherent triangle of pairwise compatibilities." This boundary relation is what makes $\partial_2 \partial_2^\top$ a non-trivial operator on C_1; without it, the curl term b_{curl} vanishes by construction and the decomposition collapses to the graph Laplacian's kernel-orthogonal split.*

B.3 The combinatorial 1-Hodge Laplacian and its kernel

The *combinatorial 1-Hodge Laplacian* of K is the symmetric positive-semi-definite operator

$$L_1 \;=\; \partial_1^\top \partial_1 \;+\; \partial_2 \partial_2^\top \;=\; L_1^{\mathrm{down}} + L_1^{\mathrm{up}} \;\in\; \mathbb{R}^{|E| \times |E|}. \tag{14}$$

Here $L_1^{\mathrm{down}} = \partial_1^\top \partial_1$ couples each edge to its endpoint-sharing neighbors (a "lower" adjacency), and $L_1^{\mathrm{up}} = \partial_2 \partial_2^\top$ couples each edge to its co-faces in T (an "upper" adjacency). The down-Laplacian L_1^{down} is the standard graph Laplacian's "edge form," while L_1^{up} is the new piece that depends on the chosen triangle set T.

The 1-Hodge Laplacian satisfies $L_1 = L_1^\top$ and $\langle L_1 b, b \rangle = \|\partial_1 b\|^2 + \left\|\partial_2^\top b\right\|^2 \geq 0$, hence is positive semi-definite. Equation 14 also implies

$$\ker(L_1) \;=\; \ker(\partial_1) \cap \ker(\partial_2^\top), \tag{15}$$

since $L_1 b = 0$ holds if and only if $\langle L_1 b, b \rangle = 0$, which forces both squared norms in the line above to vanish. The dimension of this kernel is the first Betti number of K:

$$\dim \ker(L_1) \;=\; \beta_1(K) \;=\; \dim \ker(\partial_1) - \dim \mathrm{im}(\partial_2), \tag{16}$$

which by the rank-nullity theorem applied to ∂_1 and the Euler-Poincaré formula for a 2-complex specializes to

$$\beta_1(K) \;=\; |E| - |V| + \beta_0\Big(K^{(1)}\Big) - \operatorname{rank} \partial_2, \tag{17}$$

where $\beta_0(K^{(1)}) = |\pi_0(V, E)|$ is the number of connected components of the 1-skeleton $K^{(1)} = (V, E)$. For the complete edge set $E = \binom{V}{2}$ used in this paper, $K^{(1)}$ is the complete graph K_n, so $\beta_0(K^{(1)}) = 1$ and $\beta_1(K) = \binom{n}{2} - n + 1 - \operatorname{rank} \partial_2$.

We instantiate this on the three production MoE families used throughout the paper. The per-layer candidate triangle set $T \subseteq \binom{V}{3}$ is the one constructed by Stage A of App. B.9, capped at $|T| \leq 500$; on every layer of all three models the median-pairwise-barrier subgraph supplies more than 500 qualifying 3-cliques, so the cap binds and the kept triples are a uniform-random subsample with fixed seed 42. The first Betti number $\beta_1(K) = \dim \ker(L_1)$ is a combinatorial invariant of the complex: by Eq. 17 it depends only on n and on the triangle set through $\operatorname{rank} \partial_2$, and not on the barrier values carried by the edges. Every layer of a model shares n, the complete edge set, and the triangle cap, so $\beta_1(K)$ is governed at the model level rather than by the per-layer signal, and it carries no information about where the barrier energy lies. Its role is to fix the ambient subspace $\ker(L_1)$ in which the signal-dependent quantity is measured: the harmonic energy fraction $\rho_{\mathrm{harm}}(\ell) = \left\|P_{\mathrm{harm}}\, b^{(\ell)}\right\|^2 / \left\|b^{(\ell)}\right\|^2$, the squared norm of the projection of $b^{(\ell)}$ onto $\ker(L_1)$ relative to $\left\|b^{(\ell)}\right\|^2$. A common ambient dimension is what makes ρ_{harm} comparable across layers and models, which is why Section 3.4 uses $\rho_{\mathrm{harm}}(\ell)$ as the per-layer diagnostic.

Reference orientation. The barrier signal is symmetric, $b_{ij} = b_{ji} \geq 0$, so it is embedded as a cochain relative to one fixed reference orientation: the lexicographic orientation induced by the checkpoint's expert indices, used identically for every layer and every model. The objects that depend only on K, namely $\ker L_1$, $\beta_1(K)$, the spectrum of L_1, and the weighted-operator comparison of Remark 3, are independent of this choice. The harmonic coordinates $b_{\mathrm{harm},e}$, the energy fraction ρ_{harm}, and the harmonic-critical set $E^\star$ are defined relative to the reference orientation, as edge signals on oriented simplices are in discrete Hodge theory (Lim, 2020). All reported diagnostics and selections use this one convention. The face set T is built from the pairwise-barrier graph (App. B.9), so the complex

on which P_{harm} acts is built from the pairwise barriers; the triplet-barrier values enter through the triplet-critical set $T^\star$ of Eq. 10 and the discordance fraction δ of Eq. 9. Keeping the two signals separate makes each interpretable: b_{harm} measures cyclic inconsistency of the pairwise barriers, and b_{ijk} measures joint-merge cost directly.

Remark 3 (Weighted Hodge Laplacian)**.** *A weighted variant equips vertices, edges, and triangles with inner products given by positive diagonal matrices W_0, W_E, and W_T (Horak & Jost, 2013). The adjoints of the boundary maps become $\partial_1^* = W_E^{-1}\partial_1^\top W_0$ and $\partial_2^* = W_T^{-1}\partial_2^\top W_E$, and the weighted Laplacian is $L_1^w = \partial_1^*\partial_1 + \partial_2\partial_2^*$, with harmonic subspace $H^w = \ker L_1^w = \ker\partial_1 \cap \ker(\partial_2^\top W_E)$, which does not depend on W_0 or W_T. We deliberately use the unweighted form: the edge weights b_{ij} vary by orders of magnitude across layers and across model families, and folding them into the operator would entangle "which edges are harmonic" (a topological question) with "how barriers are normalized" (a domain-specific question). The signal lives outside the operator.*

Two properties make this choice concrete. First, for a fixed complex K the projector P_{harm} does not depend on b, so for any per-layer rescaling $c > 0$,

$$P_{\mathrm{harm}}(c\,b) \;=\; c\,P_{\mathrm{harm}}(b),$$

so the normalized harmonic energy $\|P_{\mathrm{harm}}b\|^2 / \|b\|^2$ is unchanged, and so is the top-$p\%$ harmonic-critical edge set, which ranks edges by $|b_{\mathrm{harm},e}|$. Both the diagnostic and the selection are therefore invariant to per-layer barrier rescaling. Second, positive simplex weights correspond one-to-one to the cochain inner products under which the elementary cochains are orthogonal; for every such choice the kernel of the weighted Laplacian is isomorphic to the cohomology, so $\dim\ker L_1^w = \beta_1(K)$, while the harmonic projector changes in general (Horak & Jost, 2013, Sec. 2); see also Eckmann (1944). For a fixed edge inner product the triangle weights W_T cannot move the harmonic subspace, so the edge weights are the operative choice. We measured their effect on OLMoE: on all 16 layers we built L_1^w with $W_0 = I$, $W_E = \mathrm{diag}(|b_e|)/\max_e|b_e|$, and $W_T = \mathrm{diag}(|b_{ijk}|)/\max|b_{ijk}|$ (both floored at 10^{-6}), and computed the normalized projector overlap $\|P_H^w P_H\|_F^2 / \dim H$, where $H = \ker L_1$, $P_H = P_{\mathrm{harm}}$, and P_H^w is the Euclidean orthogonal projector onto H^w. The overlap has median 0.992 (minimum 0.985, maximum 0.996), the kernel dimension is unchanged on every layer, and the expected overlap of a random subspace of the same dimension is 0.72.

B.4 The discrete Hodge decomposition

We restate Theorem 1 from the main body and give a self-contained linear-algebraic proof.

Theorem 2 (Discrete Hodge decomposition)**.** *Let $K = (V, E, T)$ be a finite simplicial 2-complex with boundary operators ∂_1, ∂_2 as in App. B.2. Then $C_1(K)$ admits the orthogonal direct-sum decomposition*

$$C_1(K) \;=\; \mathrm{im}(\partial_1^\top) \,\oplus\, \mathrm{im}(\partial_2) \,\oplus\, \ker(L_1), \tag{18}$$

and consequently every $b \in C_1(K)$ has a unique decomposition $b = b_{\mathrm{grad}} + b_{\mathrm{curl}} + b_{\mathrm{harm}}$ with components in the three respective subspaces and pairwise orthogonal in the canonical inner product.

Proof. We prove three orthogonality statements and a dimension count.

Step 1 (orthogonality of $\mathrm{im}(\partial_1^\top)$ and $\mathrm{im}(\partial_2)$). Take $u = \partial_1^\top\alpha \in \mathrm{im}(\partial_1^\top)$ for some $\alpha \in C_0(K)$, and $v = \partial_2\beta \in \mathrm{im}(\partial_2)$ for some $\beta \in C_2(K)$. Then

$$\langle u, v\rangle \;=\; \langle\partial_1^\top\alpha, \partial_2\beta\rangle \;=\; \langle\alpha, \partial_1\partial_2\beta\rangle \;=\; \langle\alpha, 0\rangle \;=\; 0,$$

where the third equality is Lemma 1. Hence the two images are orthogonal.

Step 2 (orthogonality of both images to $\ker(L_1)$). If $h \in \ker(L_1)$, by Eq. 15 we have $\partial_1 h = 0$ and $\partial_2^\top h = 0$. So for any $\alpha \in C_0$ and any $\beta \in C_2$, $\langle\partial_1^\top\alpha, h\rangle = \langle\alpha, \partial_1 h\rangle = 0$ and $\langle\partial_2\beta, h\rangle = \langle\beta, \partial_2^\top h\rangle = 0$. Hence $\ker(L_1)$ is orthogonal to both images.

Step 3 (the three subspaces span $C_1(K)$). The orthogonal complement of $\mathrm{im}(\partial_1^\top)$ in C_1 is $\ker(\partial_1)$ (a standard identity in linear algebra: $\mathrm{im}(A^\top)^\perp = \ker(A)$ for any matrix A). The orthogonal complement of $\mathrm{im}(\partial_2)$ inside $\ker(\partial_1)$ is then $\ker(\partial_1) \cap \mathrm{im}(\partial_2)^\perp$, which by another application of the same identity equals $\ker(\partial_1) \cap \ker(\partial_2^\top)$. By Eq. 15, this intersection is exactly $\ker(L_1)$. Therefore

$$C_1(K) \;=\; \mathrm{im}(\partial_1^\top) \,\oplus\, \big(\ker(\partial_1)\big) \;=\; \mathrm{im}(\partial_1^\top) \,\oplus\, \big(\mathrm{im}(\partial_2) \,\oplus\, \ker(L_1)\big),$$

using the chain identity $\mathrm{im}(\partial_2) \subseteq \ker(\partial_1)$ from Lemma 1, and the orthogonal-complement identities above. Combined with Steps 1 and 2, this gives the orthogonal direct sum Eq. 18.

The decomposition $b = b_{\mathrm{grad}} + b_{\mathrm{curl}} + b_{\mathrm{harm}}$ is then unique because each summand is the orthogonal projection of b onto the corresponding subspace. □

B.5 Explicit projection formulas

The orthogonal projectors onto the three subspaces have closed-form expressions in terms of the boundary operators. Let A^+ denote the Moore-Penrose pseudoinverse of A.

Proposition 3 (Hodge projectors)**.** *The orthogonal projectors P_{grad}, P_{curl}, $P_{\mathrm{harm}} : C_1(K) \to C_1(K)$ onto $\mathrm{im}(\partial_1^\top)$, $\mathrm{im}(\partial_2)$, and $\ker(L_1)$ respectively are*

$$P_{\mathrm{grad}} \;=\; \partial_1^\top \big(\partial_1 \partial_1^\top\big)^+ \partial_1 \;=\; \partial_1^\top L_0^+ \partial_1, \tag{19}$$

$$P_{\mathrm{curl}} \;=\; \partial_2 \big(\partial_2^\top \partial_2\big)^+ \partial_2^\top \;=\; \partial_2 L_2^+ \partial_2^\top, \tag{20}$$

$$P_{\mathrm{harm}} \;=\; I_{|E|} - P_{\mathrm{grad}} - P_{\mathrm{curl}}, \tag{21}$$

where $L_0 = \partial_1 \partial_1^\top$ is the standard graph Laplacian on V and $L_2 = \partial_2^\top \partial_2$ is the analogous operator on T.

Proof. For any real matrix A, AA^+ is the orthogonal projector onto $\mathrm{im}(A)$ in the Euclidean inner product, and the singular value decomposition gives the equivalent closed form $AA^+ = A(A^\top A)^+ A^\top$ (a standard consequence of the Moore-Penrose definition; see e.g. Lim, 2020). Apply this with $A = \partial_1^\top$ to get Eq. 19, and with $A = \partial_2$ to get Eq. 20. Equation 21 follows from the orthogonal direct sum $C_1(K) = \mathrm{im}(\partial_1^\top) \oplus \mathrm{im}(\partial_2) \oplus \ker(L_1)$ and the fact that orthogonal projectors onto orthogonal subspaces sum to the identity. □

In practice, computing $P_{\mathrm{harm}} b$ does not require forming any of the three matrices L_0^+, L_2^+, P_{harm}. We instead solve two least-squares systems:

1. find $\alpha \in C_0$ minimizing $\left\|\partial_1^\top \alpha - b\right\|^2$, then set $b_{\mathrm{grad}} = \partial_1^\top \alpha$;
2. find $\beta \in C_2$ minimizing $\|\partial_2 \beta - (b - b_{\mathrm{grad}})\|^2$, then set $b_{\mathrm{curl}} = \partial_2 \beta$;
3. set $b_{\mathrm{harm}} = b - b_{\mathrm{grad}} - b_{\mathrm{curl}}$.

Each least-squares step is solved by a dense Moore–Penrose pseudoinverse (`numpy.linalg.pinv`) of $L_0 \in \mathbb{R}^{|V|\times|V|}$ for the gradient step and of $L_2 \in \mathbb{R}^{|T|\times|T|}$ for the curl step (Eqs. 19–20), giving a per-layer cost of $O\big(|V|^3 + |T|^3 + |E|\,|V|^2\big)$ with $|V| = n \le 256$ and, at the default cap, $|T| \le 500$. App. C.5 describes the evaluation order used in the timed implementation.

B.6 Harmonic energy as irreducible mergeability residual

This appendix proves Proposition 1 from main-body Section 3.3. We first make the lower-order explainable subspace and the irreducible residual energy explicit, then give the proof and an interpretation paragraph.

Recall from App. B.2 the boundary operators $\partial_1 : C_1(K) \to C_0(K)$ and $\partial_2 : C_2(K) \to C_1(K)$ satisfying $\partial_1 \partial_2 = 0$ (Lemma 1), and from Theorem 2 the orthogonal direct-sum decomposition $C_1(K) = \mathrm{im}(\partial_1^\top) \oplus \mathrm{im}(\partial_2) \oplus \ker(L_1)$ in the canonical inner product. Define the *lower-order explainable subspace*

$$\mathcal{M}_K \;:=\; \mathrm{im}(\partial_1^\top) \,\oplus\, \mathrm{im}(\partial_2) \;\subseteq\; C_1(K), \tag{22}$$

which by the same theorem is closed in $C_1(K)$. Elements of $\mathcal{M}_K$ are exactly the edge-barrier signals expressible as $m = \partial_1^\top \phi + \partial_2 \psi$ for some vertex potential $\phi \in C_0(K)$ and triangle potential $\psi \in C_2(K)$: the first term is the lift to edges of a per-vertex *is expert i broadly mergeable* score, the second term is a triangle-boundary correction that re-distributes barrier mass coherently around faces of K. The *irreducible residual energy* of $b \in C_1(K)$ relative to $\mathcal{M}_K$ is

$$\mathcal{I}_K(b) \;:=\; \inf_{\phi \in C_0(K),\, \psi \in C_2(K)} \left\|b - \partial_1^\top \phi - \partial_2 \psi\right\|^2 \;=\; \inf_{m \in \mathcal{M}_K} \|b - m\|^2, \tag{23}$$

i.e. the smallest squared error of any vertex-potential plus triangle-boundary explanation of b. Proposition 1 states $\mathcal{I}_K(b) = \|b_{\mathrm{harm}}\|^2$ with unique minimizer $m^\star = b_{\mathrm{grad}} + b_{\mathrm{curl}}$ as an element of $\mathcal{M}_K$ (the representing potentials ϕ, ψ in Eq. 23 need not be unique because $\partial_1^\top$ and ∂_2 may have non-trivial kernels).

Proof of Proposition 1. By Theorem 2, $\mathcal{M}_K$ is a closed subspace of $C_1(K)$ with orthogonal complement $\mathcal{M}_K^\perp = \ker(L_1)$. The unique decomposition $b = b_{\mathrm{grad}} + b_{\mathrm{curl}} + b_{\mathrm{harm}}$ has $b_{\mathrm{grad}} + b_{\mathrm{curl}} \in \mathcal{M}_K$ and $b_{\mathrm{harm}} \in \mathcal{M}_K^\perp$. For any $m \in \mathcal{M}_K$, write

$$b - m \;=\; \big(b_{\mathrm{grad}} + b_{\mathrm{curl}} - m\big) \;+\; b_{\mathrm{harm}},$$

where the first summand lies in $\mathcal{M}_K$ and the second in $\mathcal{M}_K^\perp$. By Pythagoras,

$$\|b - m\|^2 \;=\; \|b_{\mathrm{grad}} + b_{\mathrm{curl}} - m\|^2 \;+\; \|b_{\mathrm{harm}}\|^2 .$$

The first summand is non-negative and equals zero exactly at $m = b_{\mathrm{grad}} + b_{\mathrm{curl}}$. Therefore $\inf_{m \in \mathcal{M}_K} \|b - m\|^2 = \|b_{\mathrm{harm}}\|^2$, attained uniquely at $m^\star = b_{\mathrm{grad}} + b_{\mathrm{curl}}$. □

Interpretation. The proposition shows that the harmonic component b_{harm} is exactly the residue left in b after removing every barrier pattern expressible as (i) a vertex-level expert potential, through $\mathrm{im}(\partial_1^\top)$, and (ii) a triangle-boundary potential, through $\mathrm{im}(\partial_2)$. Therefore the normalized harmonic energy fraction

$$\rho_{\mathrm{harm}}(\ell) \;=\; \frac{\big\|P_{\mathrm{harm}}\, b^{(\ell)}\big\|^2}{\big\|b^{(\ell)}\big\|^2}$$

in Eq. 8 is the fraction of layer-ℓ edge-barrier energy that no such lower-order explanation can capture.

The statement is a property of the edge-barrier cochain b on K. Its link to the calibration KL loss $\mathcal{L}(S, \pi)$ of Eq. 1 runs through how a compression action exposes edge barriers; App. B.7 makes this link exact under an edge-exposure linearization and identifies the quantities that the HodgeCover algorithm of Section 4 acts on.

B.7 Link to compression loss under edge-exposure linearization

This appendix links the harmonic residual of Proposition 1, the part of the edge-barrier landscape that a harmonic-blind selector can miss, to the calibration KL loss $\mathcal{L}(S, \pi)$ of Eq. 1. The general corollary assumes an edge-exposure linearization of $\mathcal{L}$ (Eq. 24); for the redirect proxy of HodgeCover's own compression action, the identity holds exactly with no hypothesis (end of this subsection).

Throughout, write $a := (S, \pi)$ for a compression action and $\mathcal{L}(a) := \mathcal{L}(S, \pi)$ for its calibration KL loss (Eq. 1). Set $b_{\mathrm{lo}} := b_{\mathrm{grad}} + b_{\mathrm{curl}} \in \mathcal{M}_K$, so the Hodge decomposition reads $b = b_{\mathrm{lo}} + b_{\mathrm{harm}}$.

Edge-exposure linearization. Fix an action class $\mathcal{A}$. Suppose an ex-ante scheme fixes maps $w : \mathcal{A} \to C_1(K)$ and $\beta : \mathcal{A} \to \mathbb{R}$ (write $w_a := w(a)$, the *edge-exposure* vector, and $\beta_a := \beta(a)$) and a constant $\varepsilon \geq 0$, without using the target values $\{\mathcal{L}(a)\}_{a \in \mathcal{A}}$, such that for every $a \in \mathcal{A}$

$$\mathcal{L}(a) \;=\; \beta_a \;+\; \langle w_a,\, b\rangle \;+\; \xi_a, \qquad |\xi_a| \leq \varepsilon. \tag{24}$$

The bilinear form $\langle w_a,\, b\rangle$ aggregates how strongly action a *exposes* each pairwise merge, with β_a absorbing the zeroth-order action-dependent offset and ξ_a the linearization residual. The maps w, β and the constant ε must be fixed without the target values $\{\mathcal{L}(a)\}_{a \in \mathcal{A}}$, so that the hypothesis is stated for a scheme fixed ex ante and the offset β_a cannot encode $\mathcal{L}(a)$ itself. For any such scheme, the corollary below gives the exact form of its harmonic-blind prediction error. A harmonic-blind surrogate that sees only b_{lo} would predict

$$\mathcal{L}_{\mathrm{lo}}(a) \;:=\; \beta_a \;+\; \langle w_a,\, b_{\mathrm{lo}}\rangle. \tag{25}$$

Decompose $w_a = w_{a,\mathrm{lo}} + w_{a,\mathrm{harm}}$ along $\mathcal{M}_K \oplus \mathcal{M}_K^\perp$, with $w_{a,\mathrm{harm}} := P_{\mathrm{harm}}\, w_a$.

Corollary 1 (Harmonic exposure under edge-exposure linearization)**.** *Suppose Eq. 24 holds. For every action a,*

$$\mathcal{L}(a) - \mathcal{L}_{\mathrm{lo}}(a) \;=\; \langle w_{a,\mathrm{harm}},\, b_{\mathrm{harm}}\rangle + \xi_a, \tag{26}$$

hence

$$\big|\mathcal{L}(a) - \mathcal{L}_{\mathrm{lo}}(a)\big| \;\leq\; \|w_{a,\mathrm{harm}}\|\, \|b_{\mathrm{harm}}\| + \varepsilon. \tag{27}$$

For any two actions a, a',

$$\big(\mathcal{L}(a) - \mathcal{L}(a')\big) \;-\; \big(\mathcal{L}_{\mathrm{lo}}(a) - \mathcal{L}_{\mathrm{lo}}(a')\big) \;=\; \big\langle P_{\mathrm{harm}}(w_a - w_{a'}),\, b_{\mathrm{harm}} \big\rangle + (\xi_a - \xi_{a'}), \tag{28}$$

and consequently

$$\Big|\big(\mathcal{L}(a) - \mathcal{L}(a')\big) - \big(\mathcal{L}_{\mathrm{lo}}(a) - \mathcal{L}_{\mathrm{lo}}(a')\big)\Big| \;\le\; \|P_{\mathrm{harm}}(w_a - w_{a'})\|\, \|b_{\mathrm{harm}}\| \;+\; 2\varepsilon. \tag{29}$$

Proof. Subtracting Eq. 25 from Eq. 24 gives $\mathcal{L}(a) - \mathcal{L}_{\mathrm{lo}}(a) = \langle w_a,\, b - b_{\mathrm{lo}} \rangle + \xi_a = \langle w_a,\, b_{\mathrm{harm}} \rangle + \xi_a$. Because $b_{\mathrm{harm}} \in \ker(L_1) = \mathcal{M}_K^{\perp}$ and P_{harm} is the orthogonal projector onto $\ker(L_1)$,

$$\langle w_a,\, b_{\mathrm{harm}} \rangle \;=\; \langle P_{\mathrm{harm}}\, w_a,\, b_{\mathrm{harm}} \rangle \;+\; \langle (I - P_{\mathrm{harm}})\, w_a,\, b_{\mathrm{harm}} \rangle \;=\; \langle w_{a,\mathrm{harm}},\, b_{\mathrm{harm}} \rangle,$$

where the second term vanishes because $(I - P_{\mathrm{harm}})\, w_a \in \mathcal{M}_K$ is orthogonal to $b_{\mathrm{harm}} \in \mathcal{M}_K^{\perp}$. This proves Eq. 26; Eq. 27 is the Cauchy-Schwarz inequality applied to that inner product, plus $|\xi_a| \le \varepsilon$. Equation 28 follows by subtracting two instances of Eq. 26 and using linearity of P_{harm}, and Eq. 29 is the Cauchy-Schwarz inequality applied to the right-hand-side inner product, plus $|\xi_a - \xi_{a'}| \le 2\varepsilon$. □

Interpretation. Under the edge-exposure linearization Eq. 24, the harmonic-blind prediction error of $\mathcal{L}_{\mathrm{lo}}$ and its ranking error between any two actions are controlled exactly by the harmonic exposure of b, with no contribution from any b_{lo}-aligned exposure direction (Eqs. 26, 28). Whenever the right-hand side of Eq. 29 exceeds $|\mathcal{L}_{\mathrm{lo}}(a) - \mathcal{L}_{\mathrm{lo}}(a')|$, a harmonic-blind selector cannot certify the true ordering of $\mathcal{L}(a)$ and $\mathcal{L}(a')$. HodgeCover targets this harmonic exposure directly: it covers the top harmonic-critical edges (Eq. 10) when planning a compression, and the corollary makes precise the role harmonic-edge coverage plays under any locally-linear approximation of the calibration KL loss.

A concrete exposure vector. HodgeCover's own compression action supplies one exposure vector. For $a = (S, \pi)$ with $\pi : V \setminus S \to S$, let $e_{\{i,j\}} \in C_1(K)$ be the elementary cochain of edge $\{i, j\}$ in its reference orientation, let $c_{i\pi(i)} > 0$ be edge weights, and define

$$w_a \;:=\; \sum_{i \notin S} c_{i\pi(i)}\, e_{\{i,\pi(i)\}}, \qquad \widehat{\mathcal{L}}(a) \;:=\; \langle w_a,\, b \rangle \;=\; \sum_{i \notin S} c_{i\pi(i)}\, b_{i\pi(i)}.$$

Because each redirect edge $\{i, \pi(i)\}$ has exactly one endpoint outside S, distinct non-survivors give distinct edges, so the coefficient of every edge in w_a is well defined. The frequency weights $c_{ij} = (r_i + r_j)/r_j$ of Proposition 4 below make $\langle w_a, b^{(2)} \rangle$ the exact quadratic layer cost of an explicit model, with $b^{(2)}$ the quadratic merge barrier; the unit-weight proxy of Section 3.3 is $c \equiv 1$, and for equal positive routing frequencies $c \equiv 2$, so the quadratic model cost is then twice the unit-weight proxy and ranks plans identically. For any weights, the harmonic-blind version $\widehat{\mathcal{L}}_{\mathrm{lo}}(a) := \langle w_a, b_{\mathrm{lo}} \rangle$ satisfies, with no hypothesis,

$$\widehat{\mathcal{L}}(a) - \widehat{\mathcal{L}}_{\mathrm{lo}}(a) \;=\; \langle w_a,\, b_{\mathrm{harm}} \rangle \;=\; \langle P_{\mathrm{harm}} w_a,\, b_{\mathrm{harm}} \rangle \;=\; \sum_{i \notin S} c_{i\pi(i)}\, b_{\mathrm{harm},\{i,\pi(i)\}},$$

where the second equality holds because P_{harm} is an orthogonal projector and $b_{\mathrm{harm}} = P_{\mathrm{harm}} b$. This is Eq. 26 with $\mathcal{L} = \widehat{\mathcal{L}}$, $\beta_a = 0$, and $\xi_a = 0$: the harmonic-blind version of this proxy misprices a plan by exactly the signed harmonic exposure of its redirect edge set, and the mispricing is bounded in absolute value by $\sum_{i \notin S} c_{i\pi(i)} |b_{\mathrm{harm},\{i,\pi(i)\}}|$. The redirect of Eq. 13 weights each candidate edge by the factor $(1 + \alpha |b_{\mathrm{harm},\{i,j\}}| / \|b\|)$, traded off against the pairwise barrier b_{ij}. The identity above is exact for the proxy, and the one hypothesis that carries it to the calibration KL loss is $|\mathcal{L}(a) - \widehat{\mathcal{L}}(a)| \le \varepsilon$, which is Eq. 24 with this w_a and $\beta_a = 0$; under it, Corollary 1 applies to $\mathcal{L}$ directly, and Eq. 29 bounds the ranking error of any harmonic-blind selector by the harmonic exposure of the redirect edges plus 2ε. Proposition 4 gives a layerwise quadratic model in which the analogous identity holds with zero residual.

Exact edge-exposure model. The following setting (M1) models one MoE layer. (A1) Experts $f_1, \ldots, f_n : \mathbb{R}^{d_{\mathrm{in}}} \to \mathbb{R}^{d_{\mathrm{out}}}$ with top-1 routing on a fixed calibration multiset of N tokens; X_i is the set of tokens routed to i, $r_i = |X_i|/N$ (so $\sum_i r_i = 1$), and the layer output on $x \in X_i$ is $f_i(x)$ (a gate multiplier is absorbed into f_i). (A2) The cost of a modified layer $\tilde{y}$ is $\mathcal{C} = \frac{1}{N} \sum_x \|y(x) - \tilde{y}(x)\|^2$;

every identity below is pointwise a scalar multiple of $\|\cdot\|^2$, so it holds verbatim for any token-dependent positive semi-definite metric $\|v\|^2_{F(x)} = v^\top F(x)\, v$, including a per-token Fisher (second-order KL) metric. (A3) The redirect is a token map: tokens of X_i, $i \notin S$, are processed by $f_{\pi(i)}$, and survivor tokens are unchanged. Under a unique argmax, the max-fold of router logits, $\ell'_s = \max(\ell_s, \max_{i:\pi(i)=s} \ell_i)$ followed by $\arg\max$ over S, realizes the prescribed expert-index map: if $\arg\max \ell = i \notin S$ then $\ell'_{\pi(i)} = \ell_i$ is the global maximum, and if $\arg\max \ell = j \in S$ then $\ell'_j = \ell_j \geq \ell'_s$ for every $s \in S$. (A3) additionally stipulates a frozen (transferred) scalar gate, or no post-fold gate renormalization. (A4) Merges act in output space as in Eq. 31, $m_{ij}(x) = (r_i f_i(x) + r_j f_j(x))/(r_i + r_j)$ for $r_i + r_j > 0$, and the *quadratic merge barrier* $b^{(2)}_{ij}$ is the cost $\mathcal{C}$ of replacing both f_i and f_j by m_{ij} with routing unchanged; $b^{(2)} \in C_1(K)$ is embedded in the reference orientation like b. Write $\Delta_{ij} = f_i - f_j$ and $D^{(i)}_{ij} = \frac{1}{|X_i|}\sum_{x\in X_i} \|\Delta_{ij}(x)\|^2$ (0 if $X_i = \emptyset$). The discrepancy is *routing-symmetric* on an edge if $D^{(i)}_{ij} = D^{(j)}_{ij} =: D_{ij}$.

Proposition 4 (Exact edge-exposure model)**.** *Assume (M1).*

(i) For every plan $a = (S, \pi)$, $\mathcal{C}(a) = \sum_{i\notin S} r_i\, D^{(i)}_{i\pi(i)}$.

(ii) For $r_i + r_j > 0$, $b^{(2)}_{ij} = \dfrac{r_i r_j^2 D^{(i)}_{ij} + r_j r_i^2 D^{(j)}_{ij}}{(r_i + r_j)^2}$.

(iii) If $r_s > 0$ for every $s \in S$ and the discrepancy is routing-symmetric on every redirect edge of a, then with the directed weights $c_{ij} := (r_i + r_j)/r_j$ (i dropped, j the survivor), $c_{ij} b^{(2)}_{ij} = r_i D_{ij}$ and

$$\mathcal{C}(a) \;=\; \langle w_a,\, b^{(2)}\rangle, \qquad w_a \;=\; \sum_{i\notin S} c_{i\pi(i)}\, e_{\{i,\pi(i)\}}. \tag{30}$$

The map $a \mapsto w_a$ depends only on (S, π) and the routing frequencies. For an ex-ante scheme over an action class $\mathcal{A}$, restrict $\mathcal{A}$ to plans whose survivors have positive rates and require routing symmetry on every edge exposed by an admissible plan. With $(\mathcal{L}, b)$ replaced by the quadratic pair $(\mathcal{C}, b^{(2)})$, the algebraic conclusions of Corollary 1 then hold with zero residual: $\mathcal{C}(a) - \langle w_a, b^{(2)}_{\mathrm{lo}}\rangle = \langle P_{\mathrm{harm}} w_a, b^{(2)}_{\mathrm{harm}}\rangle = \sum_{i\notin S} c_{i\pi(i)}\, b^{(2)}_{\mathrm{harm},\{i,\pi(i)\}}$. This does not assert $\varepsilon = 0$ for the calibration KL and its KL barriers.

Proof. (i) The sets X_i partition the tokens. For $x \in X_j$ with $j \in S$ the output is unchanged. For $x \in X_i$ with $i \notin S$ the new output is $f_{\pi(i)}(x)$, contributing $\|\Delta_{i\pi(i)}(x)\|^2$. Summing and dividing by N gives $\sum_{i\notin S}(|X_i|/N)\, D^{(i)}_{i\pi(i)}$. (ii) $f_i - m_{ij} = \frac{r_j}{r_i+r_j}\Delta_{ij}$ and $f_j - m_{ij} = -\frac{r_i}{r_i+r_j}\Delta_{ij}$, and tokens outside $X_i \cup X_j$ are unaffected, so $b^{(2)}_{ij} = \frac{1}{N}\big[\sum_{X_i} \frac{r_j^2}{(r_i+r_j)^2}\|\Delta_{ij}\|^2 + \sum_{X_j} \frac{r_i^2}{(r_i+r_j)^2}\|\Delta_{ij}\|^2\big] = \frac{r_i r_j^2 D^{(i)}_{ij} + r_j r_i^2 D^{(j)}_{ij}}{(r_i+r_j)^2}$. (iii) Under symmetry, (ii) gives $b^{(2)}_{ij} = \frac{r_i r_j}{r_i+r_j} D_{ij}$, so $r_i D_{ij} = c_{ij} b^{(2)}_{ij}$; substituting into (i) gives Eq. 30. The redirect edges are distinct, since $\{i, \pi(i)\} = \{i', \pi(i')\}$ with $i, i' \notin S$ and $\pi(\cdot) \in S$ forces $i = i'$. The last display is the proof of Corollary 1 applied to $(\mathcal{C}, b^{(2)})$ with $\xi \equiv 0$, together with $\langle w, b^{(2)}_{\mathrm{harm}}\rangle = \langle P_{\mathrm{harm}} w, b^{(2)}_{\mathrm{harm}}\rangle$. □

Remark 4 (Two-sided form without symmetry)**.** *Under (M1) with $r_j > 0$, write $\mathcal{C}_i = r_i D^{(i)}_{ij}$ for a redirect edge (i dropped, $j = \pi(i)$). By Proposition 4(ii), $c_{ij}^2 b^{(2)}_{ij} = \mathcal{C}_i + \frac{r_i^2}{r_j} D^{(j)}_{ij}$, so $0 \leq \mathcal{C}_i \leq c_{ij}^2\, b^{(2)}_{ij}$, with both ends attained ($D^{(i)}_{ij} = 0$, respectively $D^{(j)}_{ij} = 0$). Suppose the discrepancy is κ-bounded, $\kappa^{-1} D^{(i)}_{ij} \leq D^{(j)}_{ij} \leq \kappa D^{(i)}_{ij}$ with $\kappa \geq 1$. If $D^{(i)}_{ij} = D^{(j)}_{ij} = 0$, then $\mathcal{C}_i = b^{(2)}_{ij} = 0$ and the bounds below hold trivially. Otherwise both discrepancies are positive, and $\mathcal{C}_i/(c_{ij} b^{(2)}_{ij}) = (r_i + r_j)/(r_j + r_i q)$ with $q = D^{(j)}_{ij}/D^{(i)}_{ij} \in [\kappa^{-1}, \kappa]$, i.e.*

$$\frac{r_i + r_j}{r_j + \kappa r_i}\, c_{ij} b^{(2)}_{ij} \;\leq\; \mathcal{C}_i \;\leq\; \frac{r_i + r_j}{r_j + r_i/\kappa}\, c_{ij} b^{(2)}_{ij},$$

and both prefactors lie in $[\kappa^{-1}, \kappa]$ (they reduce to $\kappa r_j \geq r_j$ and $r_j \leq \kappa r_j$), so $\kappa^{-1}\langle w_a, b^{(2)}\rangle \leq \mathcal{C}(a) \leq \kappa\langle w_a, b^{(2)}\rangle$: the linearization error is multiplicative, $|\xi_a| \leq (\kappa - 1)\langle w_a, b^{(2)}\rangle$.

Sources of the linearization error. In the paper's setting the residual of Eq. 24 is $\xi_a = \mathcal{L}(a) - \langle w_a, b\rangle$ with w_a from Proposition 4, and the uniform constant over an action class $\mathcal{A}$ is $\varepsilon_{\mathcal{A}} = \sup_{a\in\mathcal{A}} |\xi_a|$. Its sources are as follows. For top-k_r routing with nonnegative gates, model (M1-k) has output $y(x) = \sum_{i\in\mathcal{R}(x)} g_i(x) f_i(x)$ with $|\mathcal{R}(x)| = k_r$ and rates $r_i = |X_i|/N$ ($\sum_i r_i = k_r$), a redirect that transfers gate mass, $y_a(x) = \sum_{i\in\mathcal{R}(x)} g_i(x) f_{\bar\pi(i)}(x)$ with $\bar\pi$ the identity on S, and the output-space merge with rates r and gates unchanged, whose quadratic barrier we again write $b^{(2)}_{ij}$. Expanding the square of $y - y_a = \sum_{i\in\mathcal{R}(x)\setminus S} g_i \Delta_{i\pi(i)}$ gives exactly

$$\mathcal{C}(a) = \sum_{i\notin S} U_i + I(a), \qquad U_i = \tfrac{1}{N} \sum_{x\in X_i} g_i^2 \|\Delta_{i\pi(i)}\|^2,$$
$$I(a) = \tfrac{1}{N} \sum_x \sum_{i\neq i'\in\mathcal{R}(x)\setminus S} g_i g_{i'} \langle \Delta_{i\pi(i)}, \Delta_{i'\pi(i')}\rangle,$$

with $-\sum_{i\notin S} U_i \leq I(a) \leq (m_a - 1)\sum_{i\notin S} U_i$ for $m_a = \max_x |\mathcal{R}(x)\setminus S|$, and $I(a) = 0$ whenever no token has two dropped experts among its top-k_r. For $j = \pi(i)$ with $r_j > 0$, write $B^{(j)}_{ij} = \frac{1}{N}\sum_{X_j} g_j^2 \|\Delta_{ij}\|^2$ and the co-routing mass $Q_{ij} = \frac{1}{N}\sum_{X_i\cap X_j} g_i g_j \|\Delta_{ij}\|^2 \geq 0$. The merge changes the output on token x by $\gamma(x)\Delta_{ij}(x)$ with $\gamma = (r_j g_i \mathbb{1}_{X_i} - r_i g_j \mathbb{1}_{X_j})/(r_i + r_j)$, and averaging its square gives $(r_i + r_j)^2 b^{(2)}_{ij} = r_j^2 U_i + r_i^2 B^{(j)}_{ij} - 2 r_i r_j Q_{ij}$. Hence, with $c_{ij} = (r_i + r_j)/r_j$,

$$\mathcal{C}(a) - \sum_{i\notin S} c_{ij}\, b^{(2)}_{ij} \;=\; I(a) + \sum_{i\notin S} \Big[(c_{ij}^2 - c_{ij})\, b^{(2)}_{ij} - \Big(\frac{r_i}{r_j}\Big)^2 B^{(j)}_{ij} + 2\,\frac{r_i}{r_j}\, Q_{ij} \Big], \qquad j = \pi(i),$$

equivalently $U_i - c_{ij} b^{(2)}_{ij} = \frac{r_i}{r_i + r_j}\big(U_i - \frac{r_i}{r_j} B^{(j)}_{ij} + 2Q_{ij}\big)$. For $k_r = 1$ and $g \equiv 1$ these identities reduce to Proposition 4(i)–(ii). Accordingly, ξ_a collects: (e1) asymmetric discrepancy, the terms $(c_{ij}^2 - c_{ij}) b^{(2)}_{ij} - (r_i/r_j)^2 B^{(j)}_{ij}$ (for ungated top-1 routing, multiplicatively bounded under κ-bounded discrepancy by Remark 4); (e2) co-routing of i and $\pi(i)$, the term $2(r_i/r_j)Q_{ij}$; (e3) multi-drop interaction, $I(a)$; (e4) router modifications that are not token maps (gate renormalization or logit folds that recapture survivor tokens); (e5) the difference between the measured barrier b and the quadratic barrier $b^{(2)}$, which enters the barrier only, since the redirect cost involves no merge; (e6) the difference between the KL and its per-token second-order (Fisher) form, including cross-position terms through attention; (e7) joint compression of all layers; and (e8) rate conventions (top-k_r rates versus top-1 frequencies). Terms (e1)–(e3) have the exact expressions above.

Harmonic-blind selectors. Throughout this paragraph and Lemma 2 and Proposition 5, the model is (M1) and b denotes the quadratic barrier $b^{(2)}$ of an instance. Harmonic content is loss-relevant information in this model that no function of b_{lo}, the router, the frequencies, and the expert norms can recover. A *realizing instance* $\mathcal{I}_0 = (g, \{X_i\}, r, \{v_i\})$ of b is a layer of constant experts $f_i \equiv v_i \in \mathbb{R}^{d_{\mathrm{out}}}$, $d_{\mathrm{out}} \geq n$, with router g, routing sets X_i, frequencies $r_i > 0$, Gram matrix $\Gamma = [\langle v_i, v_j\rangle] \succ 0$, and barrier b; b is *realizable* if it has one. Its *fiber* $\mathcal{F}(\mathcal{I}_0)$ consists of constant-expert instances with the same router and routing sets, frequencies, ambient dimension, and expert norms $\|v_i\|$. Constant experts have routing-symmetric discrepancy $D^{(i)}_{ij} = \|v_i - v_j\|^2 = D^{(j)}_{ij}$, so on $\mathcal{F}(\mathcal{I}_0)$ the cost of a plan depends on the instance only through its barrier, $\mathcal{C}_b(a) = \langle w_a, b\rangle$ with w_a fixed by the frequencies (Proposition 4(iii)).

Lemma 2 (Realizability)**.** *Let $\mathcal{I}_0$ realize b. For every direction $u \in C_1(K)$ there is $t_0 > 0$ such that for all $|t| \leq t_0$ there is an instance $\mathcal{I}_t \in \mathcal{F}(\mathcal{I}_0)$ with barrier $b + tu$.*

Proof. By Proposition 4(ii), $b_{ij} = \frac{r_i r_j}{r_i + r_j}(\Gamma_{ii} + \Gamma_{jj} - 2\Gamma_{ij})$. Let $\Gamma(t)$ have the same diagonal and off-diagonal entries $\Gamma_{ij}(t) = \Gamma_{ij} - \frac{t}{2}\frac{r_i + r_j}{r_i r_j} u_{ij}$. Since $\Gamma(t)$ is continuous in t and $\Gamma(0) \succ 0$, there is $t_0 > 0$ with $\Gamma(t) \succ 0$ for $|t| \leq t_0$; then $\Gamma(t)$ is the Gram matrix of vectors $v_i(t) \in \mathbb{R}^n \subseteq \mathbb{R}^{d_{\mathrm{out}}}$ (e.g., via its Cholesky factor), with unchanged norms and barrier $b + tu$. Keeping the router and routing sets gives $\mathcal{I}_t \in \mathcal{F}(\mathcal{I}_0)$. □

A *selector* σ is a deterministic map from instances to plans with $|S| = k$; it is *blind* on a family of instances if it is constant on the family. On a fiber write $\mathrm{OPT}(b) = \min_a \mathcal{C}_b(a)$, regret $R_b(a) = \mathcal{C}_b(a) - \mathrm{OPT}(b)$, and relative excess loss $R_b(a)/\mathrm{OPT}(b)$.

Proposition 5 (Harmonic exposure lower-bounds blind regret)**.** *Assume (M1), let $\mathcal{I}_0$ realize b, and let $h \in \ker L_1$.*

(a) (Indistinguishability.) b and $b + h$ have the same b_{lo}. Whenever an instance in $\mathcal{F}(\mathcal{I}_0)$ realizes $b + h$ (by Lemma 2, for all sufficiently small multiples of any h), the two instances have the same b_{lo}, router, frequencies, and expert norms, and therefore the same scores that are functions only of those quantities, including the norm/frequency-based saliency of Eq. 10; every selector whose input is such a function is blind on the pair. These fixed summaries do not determine the harmonic perturbation: for every plan a, $\mathcal{C}_{b+h}(a) - \mathcal{C}_b(a) = \langle P_{\mathrm{harm}} w_a, h\rangle$.

(b) (Minimax lower bound.) If instances in $\mathcal{F}(\mathcal{I}_0)$ realize $b \pm h$, every selector σ blind on the pair satisfies

$$\max_{s=\pm 1} R_{b+sh}(\sigma) \;\geq\; \mathrm{Gap}(b,h) := \mathrm{OPT}(b) - \tfrac{1}{2}\big[\mathrm{OPT}(b+h) + \mathrm{OPT}(b-h)\big] \;\geq\; 0,$$

and for any two plans a_1, a_2,

$$\mathrm{Gap}(b,h) \;\geq\; \tfrac{1}{2}\Big(\big|\langle P_{\mathrm{harm}}(w_{a_1} - w_{a_2}),\, h\rangle\big| - R_b(a_1) - R_b(a_2)\Big).$$

In particular, if a_1 and a_2 are both optimal at b, the worst-case regret of every blind selector is at least half their harmonic-exposure difference, linear in the harmonic amplitude.

(c) (Explicit family.) Let $n = 3$, $k = 1$, $T = \emptyset$, $r_i = 1/3$, $z = e_{[1,2]} - e_{[1,3]} + e_{[2,3]}$ (which spans $\ker L_1$), and $b(t) = \frac{1}{6}(\mathbf{1} + tz)$ for $|t| \leq \frac{1}{2}$. All $b(t)$ are realized by unit-norm constant experts with one router, so they have the same b_{lo}, router, frequencies, and expert norms, and therefore the same scores that are functions only of those quantities. The survivor-set costs are $\mathcal{C}(\{1\}) = \mathcal{C}(\{3\}) = \frac{2}{3}$ and $\mathcal{C}(\{2\}) = \frac{2}{3}(1 + t)$, so every selector blind on the family incurs relative excess loss at least 50% at $t = \frac{1}{2}$ or $t = -\frac{1}{2}$.

Proof. (a) b_{lo} is the orthogonal projection of b onto $\mathcal{M}_K$ and $h \perp \mathcal{M}_K$, so $b + h$ has the same b_{lo}; Lemma 2 gives the realization in the fiber; and $\mathcal{C}_{b+h}(a) - \mathcal{C}_b(a) = \langle w_a, h\rangle = \langle P_{\mathrm{harm}} w_a, h\rangle$ because w_a depends only on the frequencies and $h = P_{\mathrm{harm}} h$. (b) Let a_0 be the common output of σ. Linearity gives $\mathcal{C}_{b+h}(a) + \mathcal{C}_{b-h}(a) = 2\,\mathcal{C}_b(a)$, so $\max_s R_{b+sh}(a_0) \geq \frac{1}{2}\sum_s R_{b+sh}(a_0) = \mathcal{C}_b(a_0) - \frac{1}{2}\sum_s \mathrm{OPT}(b+sh) \geq \mathrm{Gap}(b,h)$, using $\mathcal{C}_b(a_0) \geq \mathrm{OPT}(b)$. $\mathrm{Gap} \geq 0$ because OPT is a minimum of linear functions and hence concave. For the second inequality write $x_\ell = \langle w_{a_\ell}, h\rangle$ and $g_\ell = R_b(a_\ell)$. Then $\mathrm{OPT}(b+h) \leq \mathrm{OPT}(b) + g_1 + x_1$ and $\mathrm{OPT}(b-h) \leq \mathrm{OPT}(b) + g_2 - x_2$, and the same with a_1, a_2 swapped; the better pairing gives $\mathrm{OPT}(b+h) + \mathrm{OPT}(b-h) \leq 2\,\mathrm{OPT}(b) + g_1 + g_2 - |x_1 - x_2|$, and $x_1 - x_2 = \langle P_{\mathrm{harm}}(w_{a_1} - w_{a_2}), h\rangle$ since h is harmonic. (c) $\partial_1 z = 0$ and $T = \emptyset$, so $z \in \ker L_1$. With $r_i = 1/3$, $b_{ij} = \frac{1}{6} D_{ij}$, so $D = (1 + t, 1 - t, 1 + t)$ on the edges $(12, 13, 23)$. With unit norms, $\Gamma_{ij} = 1 - D_{ij}/2$, i.e. $\Gamma_{12} = \Gamma_{23} = \frac{1-t}{2}$ and $\Gamma_{13} = \frac{1+t}{2}$; the leading minors are 1, $1 - \Gamma_{12}^2 \geq \frac{7}{16}$, and $\det \Gamma(t) = \frac{1}{2} + \frac{t}{4} - t^2 + \frac{t^3}{4} \geq \frac{3}{32} > 0$ on $|t| \leq \frac{1}{2}$, so $\Gamma(t) \succ 0$ and the experts are realizable in $\mathbb{R}^3$. By Proposition 4(i), $\mathcal{C}(\{s\}) = \sum_{i \neq s} r_i D_{is}$, which gives the stated values. At $t = \frac{1}{2}$, choosing $\{2\}$ costs 1 against the optimum $\frac{2}{3}$; at $t = -\frac{1}{2}$, choosing $\{1\}$ or $\{3\}$ costs $\frac{2}{3}$ against the optimum $\frac{1}{3}$. □

Proposition 5 turns Proposition 1 from a statement about a subspace into a statement about the quadratic layer cost: on realizable constant-expert same-fiber pairs $b \pm h$, every deterministic selector constant on the pair satisfies

$$\max_{s=\pm 1} R_{b+sh}(\sigma) \;\geq\; \tfrac{1}{2}\Big(\big|\langle P_{\mathrm{harm}}(w_{a_1} - w_{a_2}), h\rangle\big| - R_b(a_1) - R_b(a_2)\Big)$$

for any two plans a_1, a_2, which is half their harmonic-exposure difference when both are optimal at b. The structure that a selector which does see b can exploit is given in App. C.2 (Proposition 6).

B.8 Effect of edge thresholding

A natural simplification is to keep only edges with $b_{ij} \leq \tau_e$ for some threshold τ_e, giving the thresholded graph $G_\tau = (V, E_\tau)$ with $E_\tau = \{(i,j) : b_{ij} \leq \tau_e\}$. We argue against this choice:

1. $\operatorname{rank} \partial_1 = |V| - |\pi_0(G_\tau)|$, so the rank of ∂_1 changes whenever edge deletions disconnect G_τ (creating a new component); deleting a non-bridge edge of G_τ leaves $\operatorname{rank} \partial_1$ unchanged but decreases $\dim \ker(\partial_1)$ by one, while deleting a bridge does the reverse;
2. the edge chain space C_1 shrinks, and every triangle $\{i,j,k\} \in T$ incident on a deleted edge is also removed, changing the rank of ∂_2 and therefore the curl subspace $\operatorname{im}(\partial_2)$;
3. $\beta_1(K)$ changes by Eq. 17, with the change depending non-monotonically on whether the deleted edge is a bridge of G_τ, lies on a non-trivial cycle, or borders a triangle in T.

The harmonic projector P_{harm} is not continuous in the combinatorial topology of K: a small change in τ_e can detach a harmonic cycle that previously contributed to $\ker(L_1)$ and re-attach it as a boundary. This makes $\rho_{\text{harm}}(\ell)$ depend strongly on τ_e rather than on the underlying barrier signal b, defeating the purpose of the diagnostic.

We therefore use $E = \binom{V}{2}$ (the complete edge set), with all barriers b_{ij} retained as the signal. Numerically, this works for the MoE sizes considered: $|E| = 2{,}016$ for OLMoE and $|E| = 32{,}640$ for both Qwen variants, and the projection uses the least-squares solves of App. B.5.

B.9 Triangle-set construction and sensitivity

The triangle set T is constructed in two stages, separating *candidate sampling* (per layer, before any Hodge analysis) from the *filtration for the Hodge complex* (per Hodge call, inside Algorithm 1).

Stage A (candidate sampling). On every MoE layer we first compute all pairwise barriers $\{b_{ij}\}_{\{i,j\} \in \binom{V}{2}}$ on the calibration corpus $\mathcal{D}$ (App. B.11). We set the candidate threshold τ_{cand} to the median of the upper-triangular entries of $\{b_{ij}\}$ and enumerate every 3-clique of the thresholded graph $(V, \{e : b_e \leq \tau_{\text{cand}}\})$, i.e. every unordered triple whose three pairwise barriers all sit at or below the median. If the number of qualifying triples exceeds the cap $|T|_{\max} = 500$ we draw a uniform-random subsample of size $|T|_{\max}$ with fixed seed 42; otherwise we keep every qualifying triple. We then compute b_{ijk} on $\mathcal{D}$ for each kept triple. The resulting $T \subseteq \binom{V}{3}$ with $|T| \leq 500$ is the layer's candidate triangle set, fixed for all subsequent Hodge calls. The cap $|T|_{\max} = 500$ is set at the high end of what one triplet-barrier sweep (one short forward pass per triple, App. C.5) finishes in the same wall-clock budget as the pairwise sweep on the largest model used here.

Stage B (filtration for the Hodge complex). Inside Algorithm 1, the Hodge complex is built at a single Betti-maximizing threshold $\tau^\star$. Concretely, we sweep τ on an 80-point uniform grid over $[0,\ 1.1 \cdot \max_e b_e]$, and at each τ build $K_\tau = (V, E_\tau, T_\tau)$ with $E_\tau = \{e : b_e \leq \tau\}$ and $T_\tau = \{(i,j,k) \in T : b_{ijk} \leq \tau$ and all three edges of (i,j,k) lie in $E_\tau\}$, recording $\beta_1(K_\tau)$ via Eq. 17. We take $\tau^\star = \arg\max_\tau \beta_1(K_\tau)$, breaking ties by the larger $|E_\tau|$. On every layer of all three models in this paper the $|E_\tau|$ tie-break drives $\tau^\star$ to the high end of the grid, where $E_{\tau^\star} = \binom{V}{2}$ and $T_{\tau^\star} = T$ (the full Stage-A candidate set); the Hodge complex used at $\tau^\star$ is therefore exactly the complete-edge / full-candidate-T complex specified in Section 3.2 and motivated in App. B.8, and the filtration sweep confirms on every layer that the complete complex maximizes β_1 and exposes the $\tau \to 0$ regime to the sensitivity discussion below. Stage B uses the candidate set T from Stage A as a fixed input; only the subset of T that survives the Stage B filtration ever enters ∂_2 and the curl subspace.

Sensitivity to $|T|$. The harmonic component b_{harm} depends on the candidate set T from Stage A through the curl subspace $\operatorname{im}(\partial_2)$, which has rank $\operatorname{rank} \partial_2 \leq \min(|E|, |T|)$. Three regimes are useful:

1. $T = \emptyset$ (no triangles): $\operatorname{im}(\partial_2) = \{0\}$ and $\ker(L_1) = \ker(\partial_1)$, the cycle space of the graph (V, E). The harmonic component then carries the entire non-gradient signal, and the decomposition reduces to the standard graph-Laplacian split into gradient and rotational components.

2. $T = \binom{V}{3}$ (all triangles): for a complete simplicial 2-skeleton on V with $E = \binom{V}{2}$, β_1 vanishes (Hatcher, 2002), so the harmonic component vanishes identically. At Qwen scale this would mean $|T| = \binom{256}{3} \approx 2.7 \times 10^6$ triangle-barrier evaluations per layer, which is computationally infeasible under the learning-free constraint.
3. T a curated subset of size $|T| = m_T \ll \binom{n}{3}$: the harmonic subspace has intermediate dimension, and *which* triples enter T is the central modeling choice. Stage A above fixes that choice (3-cliques of the median-pairwise-barrier subgraph, uniform-random-subsampled at the cap with fixed seed) and App. B.3 verifies that $|T| = 500$ holds across every layer of all three models for cross-layer and cross-model comparability.

The cap $|T|_{\max} = 500$ is a compromise between regimes 1 and 2: it keeps $|T|/\binom{n}{3}$ small enough to remain learning-free yet large enough that the kept triples concentrate on the dense, mutually low-barrier part of the barrier graph where joint merging is non-trivial to predict from any single pairwise barrier.

B.10 Frequency-weighted merge and the all-zero-frequency guard

Let $\mathcal{R}(x') \subseteq \{1, \dots, n\}$ denote the set of experts selected by the router on token x' (top-routing in all models studied here, with the model-specific fan-out taken from each backbone's released configuration; we reserve k for the survivor count and use no symbol for the router fan-out). The *empirical token-routing frequency* of expert i on the calibration corpus is the corpus-level constant

$$r_i \;=\; \mathbb{E}_{x' \sim \mathcal{D}}\, \mathbb{1}[\, i \in \mathcal{R}(x')\,] \;=\; \frac{1}{|\mathcal{D}|}\,\big|\{x' \in \mathcal{D} : i \in \mathcal{R}(x')\}\big|,$$

which depends only on $\mathcal{D}$ and not on the evaluation input x. The frequency-weighted pair merge of two experts f_i, f_j is then

$$(i \oplus j)(x) \;=\; \begin{cases} \dfrac{r_i\, f_i(x) + r_j\, f_j(x)}{r_i + r_j}, & r_i + r_j > 0, \\[2ex] \dfrac{1}{2}\big(f_i(x) + f_j(x)\big), & r_i + r_j = 0, \end{cases} \tag{31}$$

an exact convex combination on the support of the routing measure with a fallback to the unweighted average on the zero-measure exception. The implementation tests $r_i + r_j < \varepsilon$ with $\varepsilon = 10^{-12}$ in place of the exact-zero predicate to guard against floating-point underflow. An analogous formula holds for the triplet merge $(i \oplus j \oplus k)$.

The all-zero-frequency case $r_i + r_j = 0$ arises sporadically on high-expert-count Qwen layers (where some experts are routed by no calibration token), and the unguarded ratio is then $0/0$. The fallback in Eq. 31 is required to keep barrier matrices NaN-free at $n > 128$.

B.11 Calibration corpus

All barrier matrices in this paper are computed on the same calibration corpus $\mathcal{D}$: 2,048 tokens drawn from the C4 training split (Raffel et al., 2020) with random seed 42, fixed across all methods (HodgeCover, REAP, REAM, MC-SMoE, STUN, Triplet-Penalty, Triplet-Hypergraph, Greedy-Barrier), so every method sees identical calibration data.

B.12 Per-layer ranges and gradient/curl companion curves

This section reports the per-layer ranges, gradient/curl companion curves, and the representative discordance margin referenced in Figure 2.

The harmonic energy fraction $\rho_{\mathrm{harm}}(\ell)$ (Eq. 8) and the discordance fraction $\delta(\ell)$ (Eq. 9) take the following per-model ranges across $16 + 40 + 48 = 104$ layers:

- OLMoE: $\rho_{\mathrm{harm}} \in [0.289, 0.491]$ (mean 0.352); $\delta \in [0.596, 1.000]$ (mean 0.757, with $\delta = 1.000$ on layer 1, where every sampled triangle is discordant under the $1.2\times$ margin).
- Qwen3.5-35B: $\rho_{\mathrm{harm}} \in [0.387, 0.587]$ (mean 0.503); $\delta \in [0.114, 0.762]$ (mean 0.277).
- Qwen3.5-122B: $\rho_{\mathrm{harm}} \in [0.296, 0.616]$ (mean 0.456); $\delta \in [0.060, 0.892]$ (mean 0.411).

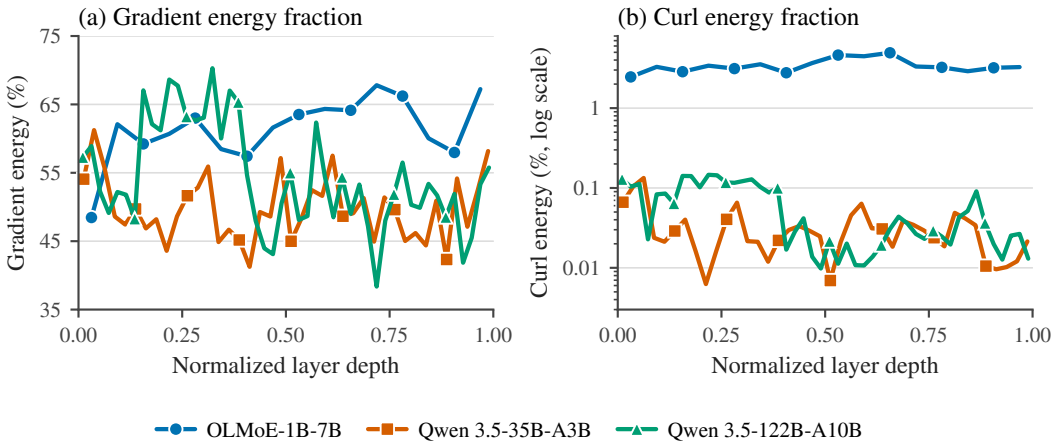

Figure 5: Gradient energy fraction $\rho_{\mathrm{grad}}(\ell)$ and curl energy fraction $\rho_{\mathrm{curl}}(\ell)$ at every layer of the same three MoE families as Figure 2. Panel (b) uses a log y-axis because OLMoE's curl fraction is roughly two orders of magnitude larger than the two Qwen variants'.

ρ_{harm} is depth-stable on every model, while δ is depth-sensitive and reveals the depth at which higher-order obstructions concentrate: the two Qwen models taper from a high value in early layers to a low value in late layers, while OLMoE remains uniformly highly discordant.

Figure 2 reports only the harmonic and discordance diagnostics. For completeness, Figure 5 reports the gradient and curl energy fractions

$$\rho_{\mathrm{grad}}(\ell) \;=\; \frac{\left\|P_{\mathrm{grad}}\, b^{(\ell)}\right\|^2}{\left\|b^{(\ell)}\right\|^2}, \qquad \rho_{\mathrm{curl}}(\ell) \;=\; \frac{\left\|P_{\mathrm{curl}}\, b^{(\ell)}\right\|^2}{\left\|b^{(\ell)}\right\|^2} \qquad (b^{(\ell)} \neq 0).$$

By Hodge orthogonality these satisfy $\rho_{\mathrm{grad}} + \rho_{\mathrm{curl}} + \rho_{\mathrm{harm}} = 1$ at every layer where $b^{(\ell)} \neq 0$. Empirically ρ_{curl} is small on the two Qwen variants ($\rho_{\mathrm{curl}} \in [6.3 \times 10^{-5}, 1.5 \times 10^{-3}]$ on Qwen3.5-35B and $\rho_{\mathrm{curl}} \in [9.8 \times 10^{-5}, 1.5 \times 10^{-3}]$ on Qwen3.5-122B at $|T| = 500$), and around 3% on OLMoE ($\rho_{\mathrm{curl}} \in [0.025, 0.049]$). The non-pairwise content of b is therefore concentrated in the harmonic component, which is the component HodgeCover covers; triangles enter the objective through the raw triplet barriers of $T^\star$, and removing them costs 3–12 pp DS-Avg on the two Qwen scales (App. F.7).

The discordance fraction δ in Eq. 9 is defined with a $1.2\times$ margin: a triangle $\{i, j, k\}$ counts as discordant if $b_{ijk} > 1.2 \cdot \max(b_{ij}, b_{ik}, b_{jk})$. We use one $1.2\times$ margin for every layer and model; δ serves as a threshold-based complement to the threshold-free ρ_{harm}, and the two agree in direction (Section 3.4).

C Algorithmic Details for HodgeCover and HodgeCover+Wanda

This appendix carries the full algorithmic specification deferred from Section 4. App. C.1 gives the per-layer HodgeCover pseudocode as Algorithm 1; App. C.3 describes the cross-layer expert-budget allocation that turns a global compression rate into the per-layer counts $\{k_\ell\}_\ell$; App. C.4 states the Stage 2 Wanda subroutine of HodgeCover+Wanda; App. C.5 records plan-time complexity and a concrete wall-clock measurement on Qwen3.5-35B (the full per-method, per-model breakdown is reported in App. F.9); App. C.6 reports the baseline-axis comparison grid (Table 3) referenced from Section 4.1.

C.1 HodgeCover pseudocode

Algorithm 1 is the complete per-layer specification of HodgeCover (Section 4.1). The triangle set T enters Algorithm 1 as a fixed input: it is the per-layer candidate set produced by Stage A of App. B.9 (3-cliques of the median-pairwise-barrier subgraph, capped at $|T| = 500$ via a fixed-seed uniform-random subsample when more triples qualify). The threshold $\tau^\star$ is the layer-specific Betti-maximizing parameter of the per-Hodge-call filtration sweep (Stage B of App. B.9, executed on line 1

Algorithm 1 HodgeCover at one MoE layer

Require: pairwise barriers $\{b_{ij}\}_{\{i,j\}\in E}$; triplet-barrier table $\{b_{ijk}\}_{(i,j,k)\in T}$; target survivor count k; non-negative saliency $\mathrm{sal} \in [0,1]^n$; protected-expert set $\mathrm{SE} \subseteq V$; hyperparameters $p, q_T, \lambda_e, \lambda_t, \alpha$

Ensure: survivor set $S^\star$ with $|S^\star| = k$, redirect map $\pi : V \setminus S^\star \to S^\star$

1: $\tau^\star \leftarrow \arg\max_\tau \ \beta_1\big(K_\tau\big)$ ▷ Betti-maximizing τ for triangle set (App. B.9)
2: Construct $K = (V, E, T)$ with the complete edge set $E = \binom{V}{2}$ and triangle set T at threshold $\tau^\star$
3: $b \leftarrow$ vector of b_{ij} on E
4: $(b_{\mathrm{grad}}, b_{\mathrm{curl}}, b_{\mathrm{harm}}) \leftarrow$ Hodge-decompose b on K (App. B.5)
5: $\|b\| \leftarrow \big(\sum_e b_e^2\big)^{1/2}$
6: $E^\star \leftarrow$ top-$p\%$ edges of E ranked by $|b_{\mathrm{harm},e}|$
7: $T^\star \leftarrow$ top-$q_T\%$ triangles of T ranked by $|b_{ijk}|$
8: $N_E[i] \leftarrow \{e \in E^\star : i \in e\}$ for all $i \in V$
9: $N_T[i] \leftarrow \{\sigma \in T^\star : i \in \sigma\}$ for all $i \in V$
10: $S \leftarrow \mathrm{SE} \cap V; \quad C_E \leftarrow \bigcup_{i\in S} N_E[i]; \quad C_T \leftarrow \bigcup_{i\in S} N_T[i]$
11: **while** $|S| < k$ **do**
12: **for** $i \in V \setminus S$ **do**
13: $\Delta(i \mid S) \leftarrow \mathrm{sal}(i) + \lambda_e \dfrac{|N_E[i] \setminus C_E|}{|E^\star|} + \lambda_t \dfrac{|N_T[i] \setminus C_T|}{|T^\star|}$
14: **end for**
15: $i^\star \leftarrow \arg\max_{i\in V\setminus S} \ \Delta(i \mid S)$
16: $S \leftarrow S \cup \{i^\star\}; \quad C_E \leftarrow C_E \cup N_E[i^\star]; \quad C_T \leftarrow C_T \cup N_T[i^\star]$
17: **end while**
18: $S^\star \leftarrow S$
19: **for** $i \in V \setminus S^\star$ **do**
20: $\pi(i) \leftarrow \arg\min_{j\in S^\star} \ b_{ij}\big(1 + \alpha\,|b_{\mathrm{harm},\{i,j\}}|/\max(\|b\|, \varepsilon_0)\big)$
21: **end for**
22: **return** $S^\star, \pi$

of Algorithm 1); the Hodge projection at $\tau^\star$ delivers both the harmonic edge mass $|b_{\mathrm{harm},e}|$ used to rank E in the top-$p\%$ step and the denominator $\|b\|$ used in the Hodge-weighted redirect cost.

A few remarks. First, the marginal-gain loop on lines 10–16 implements the standard greedy maximization of a monotone submodular function under cardinality constraint, which guarantees a $(1 - 1/e)$ approximation to the optimum of Eq. 10 (Nemhauser et al., 1978). The implementation has $O(k\,n\,(\Delta_E + \Delta_T))$ cost in marginal-gain evaluations, where Δ_E and Δ_T are the maximal incidence sizes; this is dwarfed by the barrier sweep on every layer of every model in this paper, so we adopt the plain greedy variant rather than its lazy-incremental refinements. Second, when $|E^\star| = 0$ or $|T^\star| = 0$ the corresponding normalized coverage gain is set to zero by convention; in particular, if the harmonic component vanishes (e.g., $\beta_1(K_{\tau^\star}) = 0$), the algorithm degenerates to a saliency-ranked greedy + triplet-only cover. Third, the protected set SE supports a small number of always-survive experts; in our reported runs $\mathrm{SE} = \emptyset$ on every layer of every model. Fourth, the redirect-step constant $\varepsilon_0 = 10^{-12}$ is a numerical guard against an all-zero edge-barrier signal; it is unrelated to the KL-linearization error ε in App. B.7 and to the zero-frequency guard in App. B.10.

C.2 Proof of Proposition 2 and a non-modularity example

We restate Proposition 2 for convenience: the HodgeCover objective $\Phi(S) = \sum_{i\in S} \mathrm{sal}(i) + \lambda_e\,|C_E(S)|/|E^\star| + \lambda_t\,|C_T(S)|/|T^\star|$ is non-negative monotone submodular over $S \subseteq V$ subject to $|S| = k$, and its greedy maximizer $S^\star$ satisfies $\Phi(S^\star) \geq (1 - 1/e)\ \max_{|S|=k} \Phi(S)$. The Remark 1 non-modularity example follows.

Submodularity. $\Phi(S) = \Phi_{\mathrm{sal}}(S) + \lambda_e \Phi_E(S) + \lambda_t \Phi_T(S)$ with three components. $\Phi_{\mathrm{sal}}(S) = \sum_{i\in S} \mathrm{sal}(i)$ is modular (additive) and non-negative monotone: each $\mathrm{sal}(i) \geq 0$, and adding an expert can only increase the sum. $\Phi_E(S) = |\bigcup_{i\in S} N_E(i)|/|E^\star|$ is the normalized maximum-coverage function on the universe $E^\star$ with the incidence collection $\{N_E(i)\}_{i\in V}$. Maximum-coverage functions are non-negative monotone submodular set functions: for any $A \subseteq B \subseteq V$ and any $i \notin B$, $\Phi_E(A \cup$

$\{i\})-\Phi_E(A) = |N_E(i)\backslash C_E(A)|/|E^\star| \geq |N_E(i)\backslash C_E(B)|/|E^\star| = \Phi_E(B\cup\{i\})-\Phi_E(B)$ because $C_E(A) \subseteq C_E(B)$, which is precisely the diminishing-returns characterization of submodularity (Nemhauser et al., 1978). $\Phi_T(S) = |\bigcup_{i\in S} N_T(i)|/|T^\star|$ is non-negative monotone submodular by the same argument applied to triangles. A non-negatively-weighted sum of modular and submodular functions is submodular, so Φ is non-negative monotone submodular over the ground set V with cardinality constraint $|S| = k$.

Greedy approximation. For non-negative monotone submodular Φ subject to the cardinality constraint $|S| = k$, Nemhauser et al. (1978) establish that the greedy algorithm, which iteratively selects $i_j^\star \in \arg\max_{i\in V\setminus S_{j-1}} \big(\Phi(S_{j-1}\cup\{i\}) - \Phi(S_{j-1})\big)$ starting from $S_0 = \emptyset$ until $|S_k| = k$, returns $S^\star = S_k$ satisfying

$$\Phi(S^\star) \;\geq\; \big(1-(1-1/k)^k\big)\;\max_{|S|=k}\Phi(S) \;\geq\; \big(1-1/e\big)\;\max_{|S|=k}\Phi(S), \tag{32}$$

using $(1-1/k)^k < 1/e$ for all $k \geq 1$. The greedy step in Eq. 11 (Algorithm 1, lines 10–16) is exactly this Nemhauser-Wolsey-Fisher greedy applied to Φ when $\mathrm{SE} = \emptyset$ (the empty-initialization setting used for every reported run, line 10 of Algorithm 1), so HodgeCover inherits the bound. This proves Eq. 12 of Proposition 2.

The factor $(1-1/e) \approx 0.632$ is tight in the worst case for unconstrained monotone submodular maximization under cardinality (Nemhauser et al., 1978); the guarantee we state for HodgeCover is this worst-case bound.

Non-modularity example (Remark 1). Treat the critical-edge set $E^\star$ as a given input and consider the set function $S \mapsto |C_E(S)|$. Take $V = \{1,2\}$ and $E^\star = \{\{1,2\}\}$. Adding expert 2 gains one covered edge at $S = \emptyset$ and zero after expert 1 is selected, so the marginal gain of expert 2 depends on S, whereas a modular function $\sum_{i\in S} h(i)$ gains $h(2)$ in both cases. A static score such as the incident mass $|N_E(2)| = 1$ assigns expert 2 the same value in both cases and therefore does not reproduce the greedy marginal gains of Eq. 11. □

Output-error reduction as facility location. Proposition 2 is a guarantee for the surrogate Φ that HodgeCover optimizes. Within the quadratic model of App. B.7 (M1 with routing-symmetric discrepancy), the output-error reduction of an extended action class has the same modular-plus-coverage shape. Extend the action class so that a dropped expert may be redirected or zeroed. Zeroing i costs $M_i := \frac{1}{N}\sum_{x\in X_i}\|f_i(x)\|^2$ (Proposition 4(i) with $f_{\pi(i)} \equiv 0$), and redirecting i to j costs $d_{i\to j} := \frac{1}{N}\sum_{x\in X_i}\|f_i(x) - f_j(x)\|^2$ ($= c_{ij}b_{ij}^{(2)}$ for $r_j > 0$ under routing symmetry), with $d_{i\to i} = 0$. We use $\min\emptyset = +\infty$ and $\max\emptyset = 0$.

Proposition 6 (Output-error reduction is a facility-location function)**.** *Under this model, for fixed S the optimal plan cost is $\mathcal{C}_M(S) = \sum_{i\notin S}\min\big(M_i, \min_{j\in S} d_{i\to j}\big)$; write $\mathcal{C}_M^\star := \min_{|S|=k}\mathcal{C}_M(S)$. The output-error reduction relative to zeroing the whole layer,*

$$G(S) \;:=\; \sum_i M_i - \mathcal{C}_M(S) \;=\; \underbrace{\sum_{i\in S} M_i}_{\textit{energy saliency}} \;+\; \underbrace{\sum_{i\notin S}\max_{j\in S}(M_i - d_{i\to j})^+}_{\textit{redirect coverage of dropped experts}},$$

satisfies $G(\emptyset) = 0$ and is monotone submodular; write $G^\star := \max_{|S|=k} G(S) = \sum_i M_i - \mathcal{C}_M^\star$. Greedy maximization of G returns S_g with $G(S_g) \geq (1-(1-1/k)^k)\max_{|S|=k} G \geq (1-1/e)\,G^\star$, equivalently $\mathcal{C}_M(S_g) \leq \mathcal{C}_M^\star + e^{-1}\big(\sum_i M_i - \mathcal{C}_M^\star\big)$.

Proof. By Proposition 4(i) the cost is a sum over dropped experts, each choosing its cheapest option, which gives $\mathcal{C}_M(S)$. For $i \in S$ the term of G is M_i; for $i \notin S$ it is $M_i - \min(M_i, \min_{j\in S} d_{i\to j}) = \max(0, \max_{j\in S}(M_i - d_{i\to j}))$. Hence $G(S) = \sum_i \max_{j\in S\cup\{0\}} s_{ij}$ with $s_{ii} = M_i$, $s_{ij} = (M_i - d_{i\to j})^+$ for $j \neq i$, and a phantom $s_{i0} = 0$; for $i \in S$ the maximum is s_{ii} because $s_{ij} \leq M_i$. Each $S \mapsto \max_{j\in S\cup\{0\}} s_{ij}$ with $s \geq 0$ is monotone and submodular, since the marginal gain of j, $(s_{ij} - \max_{S\cup\{0\}} s_{i\cdot})^+$, is non-increasing in S. Sums preserve both properties, and the bound of Nemhauser et al. (1978) (Eq. 32) gives the guarantee; the loss form is algebra. □

The guarantee is additive relative to the all-zero reference. The following lemma states explicit comparison conditions under which a guarantee for Φ transfers to G.

Lemma 3 (Transfer). *Let $\Phi, G \geq 0$ be set functions with Φ monotone submodular, S_Φ the greedy k-set for Φ, and $S_G^\star \in \arg\max_{|S|=k} G$. (a) If $\alpha\, G(S) \leq \Phi(S) \leq \beta\, G(S)$ for $S \in \{S_\Phi, S_G^\star\}$ with $0 < \alpha \leq \beta$, then $G(S_\Phi) \geq (1 - 1/e)\frac{\alpha}{\beta}\, G^\star$. (b) If $|\Phi(S) - (\gamma\, G(S) + \mu)| \leq \eta$ for $S \in \{S_\Phi, S_G^\star\}$ with $\gamma > 0$ and $\mu \geq 0$, then $G(S_\Phi) \geq (1 - 1/e)\, G^\star - \big(\mu/e + (2 - 1/e)\eta\big)/\gamma$.*

Proof. (a) $\beta\, G(S_\Phi) \geq \Phi(S_\Phi) \geq (1 - 1/e) \max_{|S|=k} \Phi(S) \geq (1 - 1/e)\, \Phi(S_G^\star) \geq (1 - 1/e)\, \alpha\, G^\star$. (b) $\gamma\, G(S_\Phi) + \mu + \eta \geq \Phi(S_\Phi) \geq (1 - 1/e) \max_{|S|=k} \Phi(S) \geq (1 - 1/e)(\gamma\, G^\star + \mu - \eta)$; rearrange. □

Energy-only top-k rankings can be arbitrarily suboptimal for the loss reduction:

Proposition 7 (Energy-only ranking separation). *For every $\nu \geq 2$ and every $\vartheta > 0$ there is an instance $\mathcal{I}_{\nu,\vartheta}$ of the model with constant experts, $n = \nu^2$, $k = \nu$, and equal frequencies on which every selector that takes the top k experts of a score $q(M_i)$, where q is strictly increasing, attains $G(S)/G^\star = (1 + \vartheta)^2/(\nu - 1 + (1 + \vartheta)^2)$, while greedy maximization of G is optimal. As $\vartheta \downarrow 0$ this ratio tends to $1/\nu = 1/\sqrt{n}$.*

Proof. Take ν clusters $C_1, \ldots, C_\nu$ of ν constant experts each, $v_i = \omega_c\, o_c$ for $i \in C_c$ with $\{o_c\}$ orthonormal, $\omega_1 = 1 + \vartheta$, $\omega_c = 1$ for $c \geq 2$, and $r_i = 1/n$; routing symmetry holds. The score selects $S = C_1$. For $i \in C_c$, $c \geq 2$, and $j \in C_1$, $d_{i \to j} = \frac{1}{n}(1 + (1 + \vartheta)^2) > M_i = \frac{1}{n}$, so these experts are zeroed and $G(C_1) = \nu(1 + \vartheta)^2/n$. Choosing one expert per cluster lets every dropped expert redirect at cost 0 within its cluster, so $G^\star = \sum_i M_i = (\nu(1 + \vartheta)^2 + \nu^2 - \nu)/n$, and the ratio tends to $1/\nu$ as $\vartheta \to 0$. Greedy on G gains a whole cluster at each step, so it attains $G^\star$. □

C.3 Cross-layer expert-budget allocation

Algorithm 1 is per-layer. Given a global compression rate r (as a fraction of the model's total expert count), we must turn it into per-layer survivor counts $\{k_\ell\}_\ell$ across the L MoE layers. All reported rows use the uniform allocator below; we also describe a compressibility-weighted allocator as a drop-in extension.

Uniform-per-layer allocator (HodgeCover and the REAP-family baselines). Let r be the global drop rate, n_ℓ the number of experts in MoE layer ℓ, and $R = \big\lfloor r \sum_\ell n_\ell \big\rfloor$ the total expert-drop budget. The uniform allocator sets

$$d_\ell \;=\; \big\lfloor R/L \big\rfloor + \mathbb{1}\big[\ell < (R \bmod L)\big], \qquad k_\ell \;=\; \max\big(\min\big(n_\ell - d_\ell,\; n_\ell - 1\big),\; 1\big), \tag{33}$$

so each layer drops the same number of experts up to a one-expert rounding remainder distributed deterministically to the lowest-index layers, with a per-layer floor of 1 surviving expert. This matches the published default of REAP (Lasby et al., 2025) and REAM (Jha et al., 2026), and we adopt it as the HodgeCover convention. MC-SMoE (Li et al., 2023) and STUN+Wanda (Lee et al., 2025; Sun et al., 2023) use their own native global allocators in our experiments (a global frequency-threshold rule for MC-SMoE and a global agglomerative-similarity rule for STUN+Wanda), faithful to their original papers.

Compressibility-weighted allocator (drop-in extension). A topology-aware alternative weights the drop budget by a per-layer compressibility score $\sigma(\ell) = \max\big(1 - \rho_{\mathrm{harm}}(\ell), \varepsilon_0\big)$ (layers with less harmonic mass are easier to compress), allocating

$$d_\ell \;=\; \big\lfloor R \cdot \tfrac{\sigma(\ell)}{\sum_{\ell'} \sigma(\ell')} \big\rfloor,$$

followed by the same one-expert rounding-remainder pass and the $d_\ell \leq n_\ell - 1$ floor as Eq. 33; it produces a distinct partition of the global budget on all three model scales. All HodgeCover and REAP / REAM rows in Section 5 use the uniform allocator of Eq. 33, which is the per-layer budget that Lasby et al. (2025); Jha et al. (2026) use, so that differences between these rows reflect survivor selection alone.

C.4 HodgeCover+Wanda Stage 2

Stage 2 of HodgeCover+Wanda (Section 4.2) reuses the published row-wise unstructured Wanda pruner of Sun et al. (2023) unmodified, applied independently to each surviving expert and to the shared linear layers of the MoE block (gate, up, down projections in the SwiGLU MLP, depending on

Algorithm 2 HodgeCover+Wanda Stage 2 (per surviving MLP weight matrix)

Require: weight matrix $W \in \mathbb{R}^{a\times b}$ of a survivor; calibration activations $X \in \mathbb{R}^{N\times b}$; target row sparsity $r_2 \in [0, 1)$

Ensure: pruned weight matrix $\widetilde{W}$

1: $\|X_{\cdot,j}\|_2 \leftarrow \big(\sum_{n=1}^{N} X_{n,j}^2\big)^{1/2}$ for $j = 1, \ldots, b$
2: **for** $i = 1, \ldots, a$ **do**
3: $\quad S_{i,j} \leftarrow |W_{i,j}| \cdot \|X_{\cdot,j}\|_2$ for $j = 1, \ldots, b$
4: $\quad \mathcal{K}_i \leftarrow$ indices of the top-$\lceil (1 - r_2)\, b \rceil$ entries of $\{S_{i,j}\}_{j=1}^{b}$
5: $\quad \widetilde{W}_{i,j} \leftarrow W_{i,j}$ if $j \in \mathcal{K}_i$, else $\widetilde{W}_{i,j} \leftarrow 0$
6: **end for**
7: **return** $\widetilde{W}$

architecture). For completeness, Algorithm 2 states the per-matrix loop. The single hyperparameter is the Stage-2 weight sparsity r_2, chosen so that the combined expert-axis and weight-axis compression matches the target total compression rate of the comparison cell; we never retune the Wanda saliency or the row/column choice.

The two stages share the calibration corpus $\mathcal{D}$ (App. B.11); Stage 2 reuses the activation tensor X produced during Stage 1 barrier computation, so no additional forward pass is incurred.

Numerical values of r_1 and r_2 used in our results. We fix the Stage-1 expert-axis drop rate at $r_1 = 0.20$ across every (model, rate) cell, matching the STUN-Arctic Stage-1 setting that is closest to our 64- and 256-expert backbones. Given this r_1 and a target total compression rate $r_{\text{tot}} \in \{0.33, 0.66\}$, the Stage-2 Wanda sparsity is chosen so that the combined expert-axis and weight-axis compression matches r_{tot}:

$$r_2 \;=\; \max\!\Big(0,\; \tfrac{r_{\text{tot}} - r_1}{1 - r_1}\Big) \quad\Longrightarrow\quad r_2 \;\approx\; 0.1625 \;\; (r_{\text{tot}} = 0.33), \;\; r_2 \;=\; 0.575 \;\; (r_{\text{tot}} = 0.66), \tag{34}$$

applied identically on every backbone. The matched-control hybrids REAP+Wanda and REAM+Wanda in App. F.2 use the same (r_1, r_2) schedule for parity.

C.5 Plan-time complexity and wall-clock cost

Per-layer. The asymptotically dominant term of HodgeCover at a single MoE layer is the pairwise-barrier sweep: for n experts, computing b_{ij} for every pair requires $\binom{n}{2}$ short forward passes on $\mathcal{D}$. The triplet-barrier sweep adds $|T|$ further forward passes, by construction independent of n; default and main-table runs use $|T| \le 500$, and the sensitivity panel of App. D.2 varies this cap. The Hodge projection on $K_{\tau^\star}$ is a linear-algebra operation on a complex with $|E| = \binom{n}{2}$ edges and $|T|$ triangles (App. B.5). The greedy survivor loop on lines 10–16 of Algorithm 1 is $O(k\, n\, (\Delta_E + \Delta_T))$ in marginal-gain evaluations. The sweep therefore has the per-layer bound $O\big((n^2 + |T|)\, |\mathcal{D}|\, d^2\big)$ of Table 3, and summing over the L MoE layers gives $L \cdot n^2$ pairwise barrier evaluations per model.

Implementation. The implementation behind the OLMoE and Qwen 3.5-35B-A3B plan-times of Table 14 changes the order of evaluation and leaves the algorithm unchanged. It restructures the linear algebra of the Hodge planning step by association order, for example evaluating the gradient projection as $\partial_1^\top \big(L_0^+ (\partial_1 b)\big)$ rather than as $\big(\partial_1^\top L_0^+ \partial_1\big)\, b$, and it caches projector factors that are reused within a planning run. It vectorizes the KL barrier sweep with Triton kernels that enumerate expert pairs sparsely and fuse the reductions. The optimized barrier computation agrees with the reference computation to a maximum relative difference of about 1.6×10^{-3}, and the resulting survivor plans are bit-exact with those of the reference implementation.

Measured cost. Under one uniform protocol that times the full first-run planning pipeline of every method on the same hardware, HodgeCover plans Qwen 3.5-35B-A3B in 25.0 s and HodgeCover+Wanda in 36.6 s, against 6.3 to 55.8 s across the listed baselines (Table 14). The complex and barriers are computed once per backbone and cached across compression rates (App. F.9).

Table 3: Baseline-axis comparison grid. "Survivor weights" is whether the method modifies surviving experts' weights at compression time; "Learning" is whether the method requires any gradient-based fine-tuning, knowledge distillation, or LoRA pass beyond the calibration forward passes. Plan-time complexity is per single MoE layer; n is the expert count, d the expert hidden dimension, $|\mathcal{D}|$ the calibration token count, and $|T|$ the curated triangle-set size. "Our ablation" marks baselines that we built and report only in our paper's evaluation set; they are not external published methods.

Method	Signal	Survivor weights	Plan-time per layer
HodgeCover (ours)	harm. kernel of L_1	bit-exact	$O\big((n^2 + \lvert T\rvert)\,\lvert\mathcal{D}\rvert\,d^2\big)$
HodgeCover+Wanda (ours)	harm. kernel + Wanda saliency	Wanda-pruned	$O\big((n^2 + \lvert T\rvert)\,\lvert\mathcal{D}\rvert\,d^2\big)$
REAP (Lasby et al., 2025)	gate-weighted output norm	bit-exact	$O(n\,\lvert\mathcal{D}\rvert\,d^2)^{\ddagger}$
REAM (Jha et al., 2026)	gate-coactivation + output sim.	freq-weighted merge	$O\big(n\,\lvert\mathcal{D}\rvert\,d^2 + n^2\,\lvert\mathcal{D}\rvert\big)$
MC-SMoE† (Li et al., 2023)	utilization + logit cosine	freq-weighted merge	$O\big(\lvert\mathcal{D}\rvert\,d^2 + n^2\,\lvert\mathcal{D}\rvert + n\,d^2\big)$
STUN+Wanda (Lee et al., 2025; Sun et al., 2023)	router-row sim. + Wanda saliency	cluster-mean + Wanda	$O\big(\lvert\mathcal{D}\rvert\,d^2 + n^2\,d\big)$
Greedy-Barrier (our ablation)	pairwise KL barrier	freq-weighted merge	$O\big(n^2\,\lvert\mathcal{D}\rvert\,d^2\big)$
Triplet-Penalty soft (our ablation)	pairwise + triplet KL	freq-weighted merge	$O\big((n^2 + \lvert T\rvert)\,\lvert\mathcal{D}\rvert\,d^2\big)$
Triplet-Hypergraph hard (our ablation)	pairwise KL + binary triplet veto	freq-weighted merge	$O\big((n^2 + \lvert T\rvert)\,\lvert\mathcal{D}\rvert\,d^2\big)$

†MC-SMoE in its published form (Li et al., 2023) additionally requires a 20,000-step post-merge knowledge-distillation pass against the original SMoE teacher (Section 4.1 and Table 5 of Li et al., 2023). At our 2,048-token calibration scale this KD pass overfits and regresses quality on every model in our evaluation set. We therefore report MC-SMoE without the KD step, matching the most common reduction adopted by subsequent MoE-compression literature (App. C.7).

‡The plan-time row for REAP (Lasby et al., 2025) charges the cost of the observer in the official implementation, which runs every expert on every calibration token to populate the unified pruning + merging metrics. A saliency-only implementation admits a lower $O(|\mathcal{D}|\,d^2)$ bound; we report the observer-cost form here for parity with the other baselines, all of which are charged for the dominant cost of their official reference implementations.

C.6 Baseline-axis comparison grid

Table 3 summarizes how HodgeCover differs from the seven other learning-free MoE compressors benchmarked in this paper, along three axes: which scoring signal each method assigns to experts (or expert pairs); whether surviving experts have their weights perturbed by the compression step; and plan-time complexity per single MoE layer with n experts, expert hidden dimension d, calibration corpus $\mathcal{D}$, and curated triangle set T. All table entries are learning-free as we report them; the only entry whose published form requires a learning step is MC-SMoE (Li et al., 2023), addressed in the table footnote and detailed in App. C.7. HodgeCover is the only entry that uses the harmonic kernel of the simplicial Laplacian as its scoring signal while keeping survivor weights bit-exact. The field values for each external baseline have been verified against its original published description.

Two patterns are worth flagging. First, every external baseline that perturbs survivor weights does so via a learning-free heuristic that treats the merge as a per-pair operation (Jha et al., 2026; Li et al., 2023; Lee et al., 2025); the higher-order representational risk of those merges (e.g., merging an expert that carries irreducible harmonic mass) is exactly what the harmonic-kernel coverage objective in HodgeCover was designed to sidestep, by keeping survivors bit-exact. Second, the plan-time gap between methods that score experts from a single calibration sweep (MC-SMoE, STUN+Wanda, and a saliency-only reading of REAP — leading calibration term $|\mathcal{D}|\,d^2$) and methods that assemble a pairwise-KL barrier matrix from $O(n^2)$ per-pair operations (HodgeCover, Greedy-Barrier, Triplet-Penalty, Triplet-Hypergraph — leading term $n^2\,|\mathcal{D}|\,d^2$) is a factor of up to n^2; the official REAP and REAM reference implementations additionally run every expert on every calibration token (leading term $n\,|\mathcal{D}|\,d^2$), so against them the leading terms differ by a factor of n. Measured first-run plan-time nevertheless lies within the baselines' range (Table 14), and the pipeline runs once offline, so its cost is amortized across all inference calls thereafter.

C.7 Baseline implementation notes

This subsection records implementation choices for the external baselines in Table 3 that depart from a default reading of the original paper. Three are load-bearing for the headline numbers in Section 5:

MC-SMoE without KD. Li et al. (2023) (Section 4.1, Table 5) report MC-SMoE with a mandatory 20,000-step post-merge knowledge-distillation pass against the original SMoE teacher; Table 5 of that

work shows substantial degradation when the KD pass is removed in the original paper's training-time data regime. In our offline learning-free regime the calibration corpus is $2{,}048$ C4 tokens (App. B.11); we implemented the KD module and ran it at this scale, found that it overfits the calibration corpus and degrades held-out perplexity on WikiText-103 and C4 validation by a margin larger than the gain it provides on the calibration set itself, and accordingly removed it. This matches the most common MC-SMoE reduction reported by subsequent MoE-compression literature (Lasby et al., 2025; Jha et al., 2026; Lee et al., 2025), which also benchmark MC-SMoE without the KD pass for comparability with learning-free baselines. Our MC-SMoE rows therefore report its learning-free merging stage on the shared calibration footing, the configuration used by subsequent learning-free comparisons; the KD-trained numbers of the original paper belong to a training-time regime outside this comparison.

REAP, REAM, and STUN+Wanda. We use the published reference implementations of REAP (Lasby et al., 2025), REAM (Jha et al., 2026), and STUN (Lee et al., 2025) unmodified, except that all three are run on the same fixed $2{,}048$-token C4 calibration corpus as HodgeCover (App. B.11), so all methods share identical calibration conditions. All other implementation details (cross-layer expert-budget allocator, Stage-2 Wanda saliency, etc.) follow each baseline's original implementation.

Greedy-Barrier, Triplet-Penalty soft, and Triplet-Hypergraph hard are not external baselines but ablations that we built specifically as pipeline-level comparisons on the same triplet inputs; none of the three appears in published MoE-compression literature. All three replace HodgeCover's coverage-based survivor selection with a greedy union-find merge sort, and aggregate the resulting merge groups via the same frequency-weighted average as REAM and MC-SMoE (App. B.10); they differ only in how triplet-barrier information enters the merge sort. *Greedy-Barrier* sorts all pairs $\{i, j\}$ in ascending order of pairwise barrier b_{ij} and greedily unions them via a union-find loop that terminates when exactly k components remain. No Hodge decomposition, no harmonic weighting, and no triplet input. It is the pipeline-level comparison on the pairwise-barrier signal alone. *Triplet-Penalty soft* reuses Greedy-Barrier's union-find loop with the edge cost b_{ij} replaced by $b_{ij}\,(1 + \alpha_T\,\overline{p}_{ij})$, where $\overline{p}_{ij}$ is the layer-normalized mean of triplet barriers b_{ijk} over all $(i, j, k) \in T$ that contain edge $\{i, j\}$ and $\alpha_T \geq 0$ is the penalty strength. The same triplet table T used by HodgeCover is consumed, but no Hodge decomposition is performed. *Triplet-Hypergraph hard* reuses Greedy-Barrier's union-find loop with a binary triplet veto inspired by hypergraph-cut compression (Zhou et al., 2006): a candidate union is accepted only if every triple of experts inside the resulting component (when of size ≥ 3) has $b_{ijk} \leq \tau_T$, where τ_T is the layer's 50th-percentile triplet barrier; any union that would create a component carrying a high-barrier triple is rejected and the loop continues with the next-best edge. It compares a binary triangle constraint with the soft penalty of Triplet-Penalty. All three ablations therefore use the same frequency-weighted merge aggregation as REAM/MC-SMoE; the bit-exact-survivor row of Table 3 for HodgeCover is what isolates the zero-and-redirect surgery from the topological-coverage objective. The three ablations are revisited alongside the Section 5.4 ablation table.

D Hyperparameter and Design-Choice Sensitivity

This appendix records the rationale behind, and the analytical robustness of, each scalar hyperparameter in HodgeCover. App. D.1 lists the defaults and how they were set; App. D.3 discusses the critical-simplex fractions p, q_T; App. D.4 discusses the coverage weights λ_e, λ_t; App. D.5 discusses the harmonic-redirect strength α; App. D.6 comments on the sensitivity of HodgeCover to the construction of the triangle set T; and App. D.7 comments on the sensitivity to the choice of merge metric (KL versus alternatives) used to populate the barriers b_{ij} and b_{ijk}; App. D.2 reports the measured end-to-end sensitivity panel. The hyperparameter values reported here are the values used for every model and every reported compression rate in this paper; no per-model tuning is performed. The measured panel in App. D.2 varies one knob at a time around these defaults.

D.1 Defaults and how they were set

The frozen defaults are

$$p \;=\; q_T \;=\; 20\%, \qquad \lambda_e \;=\; 1.0, \qquad \lambda_t \;=\; 0.5, \qquad \alpha \;=\; 3.0, \qquad |T| \;\leq\; 500.$$

Each value was fixed once on a single, MMLU-stratified Qwen3.5-35B layer (layer 19, the median-depth MoE layer) using a $2{,}048$-token C4-train sample disjoint from the calibration corpus $\mathcal{D}$ (App. B.11), by checking three structural criteria: whether $E^\star$ contained the edges that the per-

layer harmonic-magnitude histogram identified as "heavy"; whether the survivor set spread roughly uniformly across the harmonic-incidence-sorted ranking of experts; and whether the redirect map π avoided routing non-survivors through edges with above-mean harmonic magnitude. The same values are used for every model and rate.

D.2 Empirical sensitivity panel

This subsection measures how far the defaults of App. D.1 can move the end-to-end result. Each configuration changes one knob relative to the defaults, re-plans the compression, and is evaluated end-to-end; the panel contains 28 evaluated configurations plus a plan-level stability grid. Cells are named by backbone, rate, and axis (e.g., Qwen-35B-66 hybrid is HodgeCover+Wanda on Qwen 3.5-35B-A3B at 66%), and the default column of every table is the Table 1 value. Coverage is scored against a common reference pool per layer, formed from the union of the tested triangle sets plus a deterministic reservoir of further lowest-mean-barrier 3-cliques. Every pool triangle carries its mean pairwise barrier as a proxy for its triplet barrier, and the pool's triplet-critical triangles are the top-$q_T\%$ pool triangles under this proxy; its harmonic-critical edges are the top-$p\%$ edges by $|b_{\mathrm{harm},e}|$ on a dedicated random subsample of the pool, both at the default fractions. The pool is fixed across variants, so a change in a knob cannot change the yardstick, and it contains the tested samplers' triangles, so every variant is scored on a common pool.

Table 4: DS-Avg under one-knob changes of the triangle-set size $|T|$ (top two blocks) and the calibration size in tokens (bottom).

Cell	$\lvert T\rvert = 250$	500 (default)	1000
Qwen-35B-66 hybrid	74.58	74.6	74.45
OLMoE-66 pure	33.84	33.3	33.56
Cell	$\lvert T\rvert = 125$	500 (default)	2000
OLMoE-33 pure	46.28	46.70	45.61
Cell	2,048 (default)	4,096	8,192
OLMoE-66 hybrid	48.57	45.45	48.90
OLMoE-33 pure	46.70	44.65	46.84

Table 5: DS-Avg under one-knob changes of the critical-simplex fractions p, q_T (default 20, varied separately).

Cell	default	$p = 10$	$p = 30$	$q_T = 10$	$q_T = 30$
OLMoE-66 hybrid	48.57	48.58	48.46	48.46	48.36
OLMoE-33 pure	46.70	46.54	46.41	46.47	46.95

Table 6: DS-Avg under one-knob changes of the coverage weights (λ_e, λ_t) at the OLMoE-33 pure cell.

Cell	$(1.0, 0.5)$ (default)	$(0.5, 0.5)$	$(1, 1)$	$(2, 0.5)$
OLMoE-33 pure	46.70	46.89	47.13	46.38

Triangle set. Table 4 varies $|T|$ from 125 to 2000. Replacing the sampling seed of Stage A (App. B.9) gives 74.62 ± 0.18 (mean $\pm 2\sigma$, $n = 3$) at the Qwen-35B-66 hybrid cell and 48.45 ± 0.11 at the OLMoE-66 hybrid cell. We also replaced the sampling rule itself: a deterministic alternative sampler, which keeps the lowest-mean-barrier 3-cliques instead of a uniform-random subsample, gives 74.65 at the Qwen-35B-66 hybrid cell, against 74.63 for a re-run of the default, which reproduces Table 1 to $+0.03$ pp.

Survivor sets. The knobs change the plan little. Across the Qwen panel the survivor-set Jaccard index against the default plan is 0.999, and the plans cover 0.998 of this reference pool's harmonic-critical mass (0.997 of its proxy triplet-critical mass) under every knob setting. A separate 15-configuration

planning grid at Qwen 3.5-35B-A3B gives a median pairwise survivor-set Jaccard index of 0.977 (range 0.955 to 1.0).

Reading. Across the triangle-set, critical-fraction, and coverage-weight knobs, the observed DS-Avg ranges at non-collapsed cells span 0.22 to 1.52 pp. At the Qwen-35B-66 hybrid cell the panel also spans WikiText perplexity 10.19 to 10.33 and C4 perplexity 15.20 to 15.24. Calibration size moves DS-Avg the most: about 3 pp, non-monotone in the number of tokens, and back within 0.35 pp of Table 1 at $8{,}192$ tokens. The OLMoE-66 pure cell stays in the same regime under every variant (App. F.5).

D.3 Critical-simplex fractions p, q_T

The fractions p and q_T control the size of the harmonic-critical edge set $E^\star$ and the triplet-critical triangle set $T^\star$ that the coverage objective in Eq. 10 must cover. They have a clean problem-side interpretation: $E^\star$ is the part of K where the irreducible higher-order obstruction concentrates, and $T^\star$ is the part where joint pairwise merging amplifies pairwise cost. Two analytical properties bound how much HodgeCover can move under perturbations of p and q_T.

Saturation at large p, q_T. As $p, q_T \to 100\%$, every edge and every sampled triangle becomes critical, so $E^\star = E$ and $T^\star = T$ and the coverage gains in Eq. 11 reduce to per-expert vertex degrees in K. Since K has the complete pairwise edge set, every expert has the same vertex degree in E, so the edge-coverage gain collapses to a constant; the triangle gain remains informative through the (non-uniform) triangle incidence in T. The objective therefore degrades smoothly into a saliency-plus-triangle-degree selector at the saturation boundary.

Vanishing at small p, q_T. As $p, q_T \to 0\%$, the critical sets shrink to single edges or single triangles, and the coverage objective becomes saliency-plus-noise. Halving p or q_T from the default to 10% changes DS-Avg by a fraction of a point (Table 5).

The default $p = q_T = 20\%$ sits well inside the non-saturated, non- vanishing regime for every model in this paper: at $|E| = \binom{n}{2}$ and $n \in \{64, 256\}$, $|E^\star| \in [\,400,\ 6500\,]$, far from either boundary. The measured effect of p and q_T on DS-Avg is reported in Table 5 (App. D.2).

D.4 Coverage weights λ_e, λ_t

The coverage weights λ_e and λ_t trade off saliency against the two coverage gains in Eq. 11. Because both coverage gains are normalized by the size of the corresponding critical set, every marginal-gain term lies in $[0, 1]$ for every layer of every model; the weights λ_e, λ_t have a uniform interpretation across models. Setting $\lambda_e = \lambda_t = 0$ recovers a pure saliency-ranked selector on the REAP saliency score (Lasby et al., 2025); setting $\lambda_e \gg 1, \lambda_t \gg 1$ makes saliency a tiebreaker only. The default $\lambda_e = 1.0, \lambda_t = 0.5$ gives the edge-coverage gain parity with saliency and the triangle-coverage gain half that weight, reflecting that the triangle term is estimated from $|T| \le 500$ samples versus $\binom{n}{2}$ pairwise samples. The measured effect of (λ_e, λ_t) on DS-Avg is reported in Table 6 (App. D.2).

D.5 Harmonic-redirect strength α

The redirect cost in Eq. 13 interpolates between a pure pairwise-barrier nearest-survivor redirect ($\alpha = 0$) and a strongly harmonic-penalized redirect ($\alpha \gg 1$). The default $\alpha = 3.0$ is large enough to dominate the small differences in b_{ij} that occur for nearby survivors but small enough not to override large pairwise-barrier gaps: at the default, two survivors with equal b_{ij} to a non-survivor break the tie according to the harmonic component, and $\alpha = 3$ corresponds to a $4\times$ penalty on the most-harmonic edge in the layer relative to a zero-harmonic edge (i.e., $1 + 3 \cdot 1 = 4$ versus 1).

D.6 Triangle-set construction

The sensitivity of b_{harm} to the choice of triangle set T is analyzed in App. B.9. There we contrast three regimes: $T = \emptyset$ (pairwise-only; the curl subspace $\mathrm{im}(\partial_2)$ vanishes and the harmonic component reduces to the graph cycle space $\ker(\partial_1)$, which then carries the entire non-gradient signal), $T = \binom{V}{3}$ (every triple a face; combinatorially intractable at $n = 256$ where $|\binom{V}{3}| > 2.7 \times 10^6$), and the curated $|T| \le 500$ used here. The third regime is what all results in this paper use, with $|T| \le 500$ in default and main-table runs and the cap varied in the sensitivity panel; the cap of 500 was chosen at the high end of what the triplet-barrier sweep (one forward pass per triple) can finish in the same wall-clock

budget as the pairwise sweep. The measured effect of $|T|$, of the sampling seed, and of a deterministic alternative sampling rule is reported in App. D.2.

D.7 Merge metric

All barriers b_{ij}, b_{ijk} in this paper are computed from the KL divergence between the original layer's output distribution and the merged layer's output distribution on the calibration corpus $\mathcal{D}$ (App. B.10). KL is the natural choice in the language-modeling regime: the per-token output is a categorical distribution over the vocabulary, and the calibration objective in Eq. 1 is itself a KL. Other f-divergences (Jensen- Shannon, Hellinger), or distributional metrics on activations rather than outputs (cosine on activated-output vectors, Wasserstein-1 on activation distributions), would all populate b_{ij}, b_{ijk} with positive scalars and could be substituted, since the Hodge construction is agnostic to the metric; we use KL throughout because Eq. 1 is a KL. The chain-complex structure of K and the Hodge decomposition itself depend only on which pairs are in E and which triples are in T, not on the metric used to populate the edge-supported signal b, so a metric swap is local to the barrier sweep.

E Experimental setup details

This appendix supports Section 5.1.

E.1 Models and checkpoints

We benchmark three open-weight sparse MoE backbones, all pulled from their canonical releases without fine-tuning. OLMoE-1B-7B (`allenai/OLMoE-1B-7B-0924`, 16 MoE layers, 64 experts per layer, top-2 routing) instantiates the small-model regime. Qwen 3.5-35B-A3B (40 MoE layers, 256 experts per layer, top-8 routing) and Qwen 3.5-122B-A10B (48 MoE layers, 256 experts per layer, top-8 routing) instantiate the high-cardinality regime; we use the base checkpoints with reinforcement-learning post-training disabled to keep the calibration distribution and the baseline router both well-defined. Tokenizer, vocabulary, and positional encoding follow each backbone's default configuration.

E.2 Compression rates and the per-layer allocator

Compression rates are quoted as fractions of total expert count across the model: 33% removes $\lfloor 0.33 \cdot \sum_\ell n_\ell \rfloor$ experts in total and 66% removes $\lfloor 0.66 \cdot \sum_\ell n_\ell \rfloor$. The per-layer survivor counts $\{k_\ell\}$ that realise this total are produced by each baseline's source-paper allocator (App. F.1). We chose 33% and 66% rather than 25%/50%/75% because the two rates straddle the "easy / hard" frontier reported by the prior MoE-compression literature (REAP, REAM, and STUN+W all report 50%-rate numbers in the modest-loss regime; 66% is past the regime in which the published hybrids retain their downstream profile).

E.3 Calibration and evaluation

Calibration uses 2,048 tokens drawn from the C4 train split (Raffel et al., 2020) at random seed 42, fixed across all methods. Tokens are concatenated to the model's training context length and truncated to fit the calibration budget; we did not subsample by document, since the relevant signal for the per-layer barriers is distributional rather than per-document. Evaluation reports WikiText-103 perplexity (Merity et al., 2016), C4 perplexity (held-out validation split, never overlapping with calibration), and nine downstream tasks via the LM Evaluation Harness (Gao et al., 2021) at version `v0.4.x`: ARC-Challenge and ARC-Easy (Clark et al., 2018) (zero-shot, accuracy-norm), BoolQ (Clark et al., 2019) (zero-shot), HellaSwag (Zellers et al., 2019) (zero-shot, accuracy-norm), MMLU (Hendrycks et al., 2020) (5-shot, macro across the 57 subjects), PIQA (Bisk et al., 2020) (zero-shot), TruthfulQA-MC2 (Lin et al., 2022) (zero-shot), WinoGrande (Sakaguchi et al., 2021) (zero-shot), and GSM8K (Cobbe et al., 2021) (8-shot, flexible-extract grading). DS-Avg in this paper is the unweighted macro of the nine downstream accuracies.

E.4 Hardware and reproducibility

Calibration, plan-time, and topology audits run on a single workstation with 2 NVIDIA RTX PRO 6000 Blackwell GPUs (98 GB each, 252 GB RAM, 64 CPU cores) for the primary runs; the OLMoE benchmarks were run on a paired A6000 workstation, and the Qwen 3.5 ablations and topology audits were distributed across H200 nodes for parallel throughput. Inference throughput is measured with

PyTorch 2.x and Hugging Face `transformers` at the model's default context length, batch size 1, 128-token decode, repeated 5 times with the median reported.

F Additional Results

This appendix supports Section 5. It is organized as nine subsections: a description of the per-method cross-layer allocator and the four HodgeCover ablations (App. F.1), the full hybrid axis with matched-control hybrids and a per-rate Pareto plot (App. F.2), statistical uncertainty of the headline comparisons (App. F.3), an additive per-expert selector baseline (App. F.4), an analysis of the OLMoE 66% expert-count cell (App. F.5), the per-layer retained mass trajectories for the harmonic / gradient / curl / triplet-barrier components on the two backbones not shown in the main body (App. F.6), the per-task ablation breakdown across all three backbones (App. F.7), a vision-language study (App. F.8), and the plan-time and throughput (App. F.9).

F.1 Per-method allocator and ablation construction

Per-method baseline implementation. For each baseline we follow the source paper's released implementation; the brief notes below summarize the survivor-selection rule used in our experiments.

REAP (Lasby et al., 2025) ranks experts within each MoE layer by an output-saliency score and drops the lowest-ranking ones, applying the uniform per-layer cut of Eq. 33.

REAM (Jha et al., 2026) reduces every MoE layer to a fixed target survivor count by saliency-guided clustering of experts and merging non-centroid members into their nearest centroid.

MC-SMoE (Li et al., 2023) uses normalized routing frequencies to mark a global pool of low-frequency experts as non-dominant subject to a per-layer "at-least-one-dominant" floor, then merges each non-dominant expert into the most router-similar dominant expert in its layer.

STUN+Wanda (Lee et al., 2025; Sun et al., 2023) composes the behavioral-similarity Stage-1 of STUN with a Stage-2 unstructured Wanda (Sun et al., 2023) sweep on survivor weights, following the released STUN implementation; the Stage-2 protocol is in App. C.4.

HodgeCover (ours) follows the REAP convention with the uniform per-layer allocator of Eq. 33.

The four ablations of HodgeCover. The four ablations probe two orthogonal levers of the HodgeCover construction: whether the triangle term contributes beyond the harmonic-edge term (*Hodge No-Triangle*), and how the full HodgeCover pipeline compares with merge-based pipelines that consume the same input data without the Hodge decomposition (*Greedy-Barrier*, *Triplet-Penalty soft*, *Triplet-Hypergraph hard*). Hodge No-Triangle reuses Algorithm 1's backbone unchanged and only modifies the coverage objective; the other three ablations replace survivor selection with a greedy union-find merge sort and aggregate the resulting merge groups via the same frequency-weighted average as REAM and MC-SMoE (App. B.10).

Hodge No-Triangle. Set $\lambda_t = 0$ in Eq. 10, dropping the triplet-triangle coverage term entirely and reducing Φ to saliency plus harmonic-edge coverage. The Step 1 harmonic-critical edge set $E^\star$ is still extracted from the Hodge decomposition; only the triangle term is removed. This isolates the contribution of the triangle term holding the harmonic-edge term fixed.

Greedy-Barrier (no topology). Sort all pairs $\{i, j\}$ in ascending order of pairwise barrier b_{ij} and greedily union them via a union-find loop that terminates when exactly k components remain. No Hodge decomposition, no harmonic weighting, and no triplet input. This is the pipeline-level comparison against the cleanest topology-blind merge.

Triplet-Penalty soft (no Hodge decomposition). Reuse Greedy-Barrier's union-find loop with the edge cost b_{ij} replaced by $b_{ij}\,(1 + \alpha_T\,\overline{p}_{ij})$, where $\overline{p}_{ij}$ is the layer-normalized mean of triplet barriers b_{ijk} over all $(i, j, k) \in T$ that contain edge $\{i, j\}$ and $\alpha_T \geq 0$ is the penalty strength. The same triplet table T that HodgeCover consumes is consumed here, but the Hodge decomposition is skipped.

Triplet-Hypergraph hard (binary triangle veto). Reuse Greedy-Barrier's union-find loop with a hard triplet veto inspired by hypergraph-cut compression (Zhou et al., 2006): a candidate union is accepted only if every triple of experts inside the resulting component (when of size ≥ 3) has $b_{ijk} \leq \tau_T$, where τ_T is the layer's 50th-percentile triplet barrier; any union that would create a component carrying a

high-barrier triple is rejected and the loop continues with the next-best edge. This is the hard-veto analogue of Triplet-Penalty soft and compares a binary triangle constraint with a soft penalty.

The four ablations span triangle term on/off and Hodge decomposition on/off (HodgeCover on/on, Hodge No-Triangle off/on, Triplet-Penalty soft on/off, Greedy-Barrier off/off; Triplet-Hypergraph hard is the hard-constraint variant of on/off). Hodge No-Triangle holds the pipeline fixed; the three merge-based ablations compare full pipelines.

F.2 Full hybrid axis with matched-control baselines

Section 5.2 reports STUN+Wanda as the published hybrid baseline. The matched-control hybrids REAP+Wanda and REAM+Wanda are not reported in their source papers; we run them as matched controls under the same Stage-2 protocol. The Stage 2 residual-sparsity protocol of App. C.4 is applied identically across all four hybrid methods. Table 7 reports DS-Avg and PPL for each (model, rate) cell.

Table 7: Full hybrid axis on all three backbones, including the matched-control hybrids REAP+Wanda and REAM+Wanda (computed by us; not reported in the source REAP / REAM papers). DS-Avg is the unweighted macro of the nine downstream tasks.

Method	Wiki	C4	DS-Avg
OLMoE-1B-7B — 33% *rate*			
STUN+Wanda (Lee et al., 2025; Sun et al., 2023)	**11.85**	18.43	50.8
REAP+Wanda (Lasby et al., 2025; Sun et al., 2023)	14.18	16.98	52.9
REAM+Wanda (Jha et al., 2026; Sun et al., 2023)	14.32	16.87	47.0
HodgeCover+Wanda (ours)	13.92	**16.69**	**53.0**
OLMoE-1B-7B — 66% *rate*			
STUN+Wanda (Lee et al., 2025; Sun et al., 2023)	22.11	31.01	45.1
REAP+Wanda (Lasby et al., 2025; Sun et al., 2023)	22.83	24.91	43.3
REAM+Wanda (Jha et al., 2026; Sun et al., 2023)	31.65	26.43	40.8
HodgeCover+Wanda (ours)	**18.32**	**21.18**	**48.6**
Qwen 3.5-35B-A3B — 33% *rate*			
STUN+Wanda (Lee et al., 2025; Sun et al., 2023)	**8.86**	15.61	66.0
REAP+Wanda (Lasby et al., 2025; Sun et al., 2023)	9.01	13.63	**76.3**
REAM+Wanda (Jha et al., 2026; Sun et al., 2023)	9.01	13.59	75.8
HodgeCover+Wanda (ours)	9.18	**13.42**	76.1
Qwen 3.5-35B-A3B — 66% *rate*			
STUN+Wanda (Lee et al., 2025; Sun et al., 2023)	11.77	20.05	62.0
REAP+Wanda (Lasby et al., 2025; Sun et al., 2023)	10.57	15.78	74.1
REAM+Wanda (Jha et al., 2026; Sun et al., 2023)	10.48	15.68	73.8
HodgeCover+Wanda (ours)	**10.25**	**15.22**	**74.6**
Qwen 3.5-122B-A10B — 33% *rate*			
STUN+Wanda (Lee et al., 2025; Sun et al., 2023)	6.35	14.48	75.5
REAP+Wanda (Lasby et al., 2025; Sun et al., 2023)	6.13	12.87	77.1
REAM+Wanda (Jha et al., 2026; Sun et al., 2023)	**5.98**	12.70	**77.9**
HodgeCover+Wanda (ours)	5.99	**12.49**	77.8
Qwen 3.5-122B-A10B — 66% *rate*			
STUN+Wanda (Lee et al., 2025; Sun et al., 2023)	8.77	17.50	70.8
REAP+Wanda (Lasby et al., 2025; Sun et al., 2023)	7.98	14.52	75.4
REAM+Wanda (Jha et al., 2026; Sun et al., 2023)	7.78	14.25	75.6
HodgeCover+Wanda (ours)	**7.42**	**13.86**	**75.9**

At 33% on the two Qwen scales the four hybrids lie within 0.8 pp DS-Avg of each other. At 66% HodgeCover+W leads REAP+W and REAM+W by 0.3–0.8 pp and STUN+W by 5–13 pp DS-Avg on both Qwen scales, and on OLMoE-1B-7B it leads all three (HodgeCover+W 48.6% vs. STUN+W 45.1%, REAP+W 43.3%, REAM+W 40.8%). The lead therefore holds against the strongest matched hybrids, not only against STUN+W.

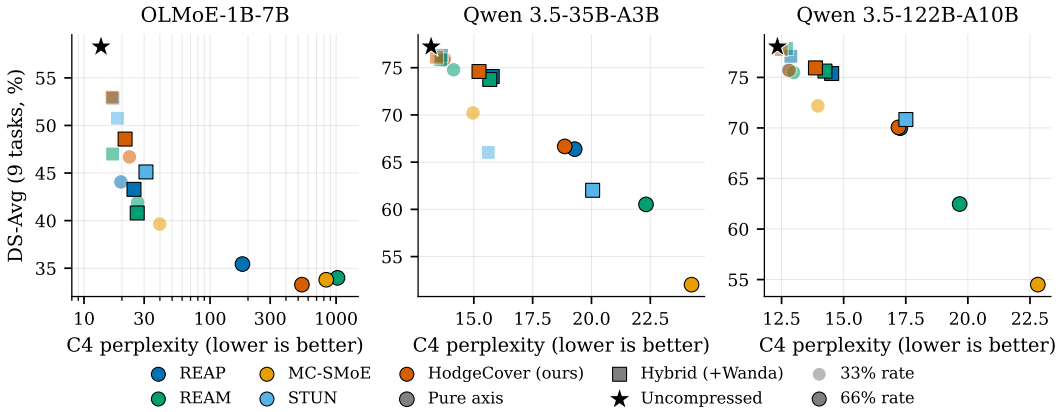

Figure 6: C4 perplexity vs. DS-Avg at 33% (lighter, white-edged) and 66% (darker, black-edged) rates. Pure expert-count axis = circles, hybrid axis = squares. Uncompressed reference is the black star. The OLMoE panel uses a logarithmic x axis. HodgeCover+Wanda sits on the upper-left frontier of every panel.

F.3 Statistical uncertainty of the headline comparisons

Table 1 reports single-run point estimates. This subsection attaches three kinds of uncertainty to them: evaluation noise for a fixed compressed model, paired tests between methods, and replicates of the full compression across calibration seeds.

Benchmark-sampling noise. The LM Evaluation Harness reports a standard error for each task score. We propagate these standard errors through the nine-task macro to obtain an approximate 95% interval for DS-Avg across the Table 1 cells with per-task standard errors, on all three backbones. The macro-level standard errors span 0.32 to 0.37 pp. These intervals describe evaluation noise for a fixed compressed model and are unpaired across methods.

Paired tests. Two methods evaluated on the same examples share example-level variance, which unpaired intervals do not remove. We therefore ran paired per-example bootstrap and permutation tests (10k permutations) on the per-example evaluation logs, with a Holm correction over the five-contrast family of Table 8. The resampling unit is the evaluation example, stratified by task, and each resample is macro-aggregated over the nine tasks as in DS-Avg. The paired deltas are computed from the per-example logs of matched re-runs of the two arms and therefore differ from the Table 1 point-estimate gaps; in particular, both OLMoE contrasts give +5.88.

Table 8: DS-Avg contrasts at 66% (pp). Unpaired: difference of point estimates with the propagated benchmark-sampling 95% interval. Paired: per-example paired difference with the Holm-corrected permutation p-value (10k permutations, five-contrast family).

Contrast (66%)	Unpaired Δ [95% CI]	Paired Δ, Holm p
OLMoE HodgeCover+W vs. REAP+W	+5.28 [+4.37, +6.20]	+5.88, $p = 5 \times 10^{-4}$
OLMoE HodgeCover+W vs. STUN+W	+3.46 [+2.53, +4.38]	+5.88, $p = 5 \times 10^{-4}$
Qwen-35B HodgeCover+W vs. REAP+W	+0.53 [−0.39, +1.45]	+0.51, $p = 9.1 \times 10^{-3}$
Qwen-35B HodgeCover+W vs. REAM+W	+0.85 [−0.07, +1.77]	+0.57, $p = 3 \times 10^{-3}$
Qwen-35B HodgeCover vs. REAP (pure)	+0.29 [−0.58, +1.16]	+0.16, $p = 4 \times 10^{-2}$

All five contrasts reject the null of no difference, with HodgeCover ahead. The unpaired intervals cross zero at the three Qwen 3.5-35B-A3B cells; the paired tests resolve these margins because pairing removes the example-level variance the two methods share. At 33% on the pure Qwen 3.5-35B-A3B axis, the gap between HodgeCover and REAP (75.9 vs. 75.8) is 0.1 pp, inside the benchmark-sampling noise above, and we treat it as a tie.

Calibration-seed replicates. On OLMoE-1B-7B we repeated the full compression with calibration seeds 42, 43, and 44; the seed drives survivor selection, saliency, and the Wanda inputs. The

HodgeCover margin over REAP is positive at every seed on both axes: $+2.44$, $+0.75$, and $+2.27$ pp at 33% on the pure axis, and $+5.88$, $+1.90$, and $+6.53$ pp at 66% on the hybrid axis (HodgeCover+W vs. REAP+W; $n = 3$, hybrid sample standard deviation 2.5 pp).

Scope. Paired tests cover OLMoE-1B-7B and Qwen 3.5-35B-A3B, seed replicates cover OLMoE-1B-7B, and benchmark-sampling intervals cover all three backbones.

F.4 Additive per-expert selector baseline

HodgeCover selects survivors by greedy coverage of the harmonic-critical edge set $E^\star$ and the triplet-critical triangle set $T^\star$. A simpler selector could consume the same quantities as static per-expert scores. We implemented such a selector as a selector-only control: the cross-layer allocator, the redirect path, the survivor treatment, the Wanda stage, the calibration corpus, and the per-layer budgets are frozen to HodgeCover's, and only the selection rule changes. The additive selector scores expert i by

$$h(i) = \mathrm{sal}(i) + \lambda_e \frac{\sum_{e \in E^\star : i \in e} |b_{\mathrm{harm},e}|}{\sum_{e \in E^\star} |b_{\mathrm{harm},e}|} + \lambda_t \frac{\sum_{\sigma \in T^\star : i \in \sigma} |b_\sigma|}{\sum_{\sigma \in T^\star} |b_\sigma|},$$

with the saliencies $\mathrm{sal}(i) \in [0, 1]$, the sets $E^\star$, $T^\star$, and the triplet barriers b_σ of HodgeCover (a ratio is set to 0 when its denominator vanishes), and takes the top-k experts per layer. Table 9 reports DS-Avg at the fixed default weights $(\lambda_e, \lambda_t) = (1.0, 0.5)$ of HodgeCover; the OLMoE 66% hybrid cell was additionally run at two other weight settings as a sensitivity check.

Table 9: DS-Avg of HodgeCover and of the additive per-expert selector under an otherwise identical pipeline. In hybrid cells both selectors are followed by the same Wanda stage.

Cell	HodgeCover	Additive
Qwen-35B 33% pure	75.95	75.74
Qwen-35B 66% pure	66.67	65.95
Qwen-35B 66% hybrid	74.63	74.67
OLMoE 33% pure	46.70	46.57
OLMoE 66% hybrid	48.57	46.68
Qwen-122B 66% pure	70.06	69.51
Qwen-122B 66% hybrid	75.94	76.10

HodgeCover is higher in five of the seven cells, by $+0.13$ to $+1.89$ pp, and lower by 0.04 and 0.16 pp in the other two. At the Qwen-35B 66% hybrid cell both selectors stay above the REAP+W, REAM+W, and STUN+W hybrids (App. F.2). The additive selector reproduces HodgeCover's survivor sets at Jaccard index 0.89 to 0.98, confirming that the harmonic-critical and triplet-critical quantities carry the selection signal; greedy coverage optimizes them directly with the $(1 - 1/e)$ guarantee of Proposition 2, while the additive weight grid alone moves OLMoE by 2.65 pp.

F.5 The OLMoE 66% expert-count cell

At 66% on the pure expert-count axis of OLMoE-1B-7B, every learning-free expert-count method crosses C4 perplexity 100 (Table 1).

Minimum hitting set. For each layer we solve the integer program

$$\min_{x \in \{0,1\}^n} \sum_{i \in V} x_i \quad \text{s.t.} \quad \sum_{i \in \sigma} x_i \;\geq\; 1 \quad \text{for every } \sigma \in E^\star \cup T^\star,$$

whose value is the minimum number of experts that touch every harmonic-critical edge and every triplet-critical triangle. The value depends only on $E^\star$ and $T^\star$, so it lower-bounds the budget that any selection rule needs to cover these critical sets fully. At 66%, every layer's allocated survivor count is below the layerwise minimum (34 to 40), whereas every layer's 33% allocation clears that minimum, which is consistent with the lower critical-mass coverage of the plans at this cell: they cover less of the frozen reference pool's critical mass (App. D.2; harmonic 0.80, triplet 0.57) than the plans at the other OLMoE cells (at least 0.96).

Triangle-set variants. This regime is not an artifact of the sampled triangle set. Table 10 re-plans the cell under the triangle-set variants of App. D.2, and every variant stays in the same regime (DS-Avg 33.24–34.20).

Table 10: HodgeCover at the OLMoE 66% pure cell under triangle-set variants: triangle-set size, Stage-A sampling seed, and the deterministic alternative sampling rule.

	default	$\|T\| = 250$	$\|T\| = 1000$	seed 1	seed 2	alt. rule
DS-Avg	33.3	33.84	33.56	33.53	33.24	34.20
WikiText PPL	1,134.9	928	1,049	1,049	1,325	1,067

At the same 66% total-compression target, the hybrid axis remains usable on this backbone: with Wanda on the survivors, HodgeCover+Wanda posts its largest margin over the matched REAP+Wanda control (+5.3 pp; App. F.2), and the paired test of this margin is in App. F.3.

F.6 Mechanism: H/G/C/T retained mass on OLMoE and Qwen 3.5-35B-A3B

Retained-mass metric. Fix MoE layer ℓ with simplicial mergeability complex $K^{(\ell)} = (V^{(\ell)}, E^{(\ell)}, T^{(\ell)})$ and edge-supported barrier signal $b^{(\ell)} \in C_1(K^{(\ell)})$ on the *uncompressed* layer (Section 3). The Hodge decomposition (Theorem 1) splits $b^{(\ell)}$ into gradient, curl, and harmonic components $b^{(\ell)} = b^{(\ell)}_{\text{grad}} + b^{(\ell)}_{\text{curl}} + b^{(\ell)}_{\text{harm}}$; write $h^{(\ell)}_e$, $g^{(\ell)}_e$, $c^{(\ell)}_e$ for the per-edge coefficients of the harmonic, gradient, and curl components, and $b^{(\ell)}_\sigma$ for the triplet-barrier value on triangle $\sigma \in T^{(\ell)}$. Given a survivor set $S^{(\ell)} \subseteq V^{(\ell)}$ produced by any compression method, the per-layer retained-mass metrics are the fractions of ℓ^1 mass on edges and triangles whose support *intersects* the survivor set:

$$\text{ret}^{(\ell)}_X(S) \;=\; \frac{\sum_{e \in E^{(\ell)}:\, e \cap S \neq \emptyset} |X^{(\ell)}_e|}{\sum_{e \in E^{(\ell)}} |X^{(\ell)}_e|}, \qquad X \in \{\text{harm, grad, curl}\}, \tag{35}$$

and analogously for the triplet-barrier component, replacing the edge sum by a sum over triangles $\sigma \in T^{(\ell)}$ with $\sigma \cap S \neq \emptyset$, weighted by $|b^{(\ell)}_\sigma|$. Survivors are evaluated against the *original* (pre-compression) Hodge decomposition rather than re-decomposing the survivor sub-MoE's complex, so the four metrics directly measure how much of the original layer's structural signal each method's survivor set covers. The macro retained-mass at a given (model, rate) cell is the unweighted average of $\text{ret}^{(\ell)}_X(S^{(\ell)})$ over MoE layers.

Section 5.3 plots the four-component retained mass on Qwen 3.5-122B-A10B at 66%. Figures 7 and 8 extend the same plot to the two remaining backbones.

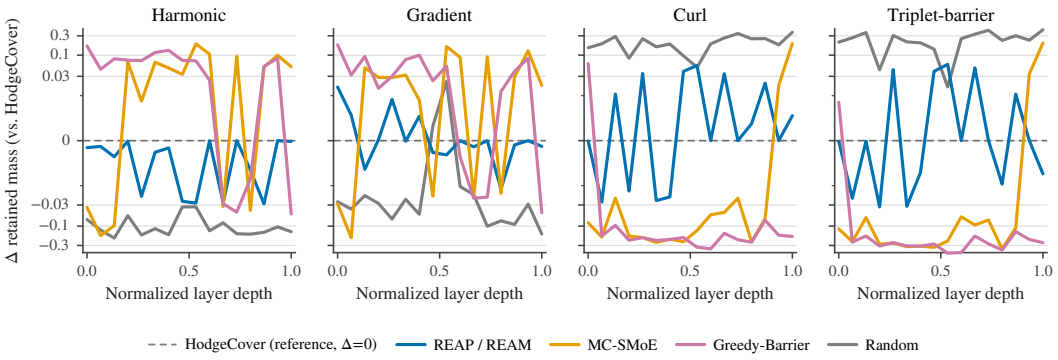

Figure 7: Per-layer deviation from HodgeCover on OLMoE-1B-7B at 66% across the four Hodge components.

The *H/G/C/T* trade-off observed in Section 5.3 is consistent across the two remaining backbones. Greedy-Barrier and MC-SMoE retain more harmonic and gradient mass than HodgeCover at the cost of curl and triplet-barrier mass; Random sweeps in the opposite direction; REAP and REAM share identical pure-axis survivor selection (their plots are drawn as a single curve in Figure 4) and deviate from HodgeCover weakly on every component. Table 11 reports the macro retained mass at 66% for the same five methods across the four components.

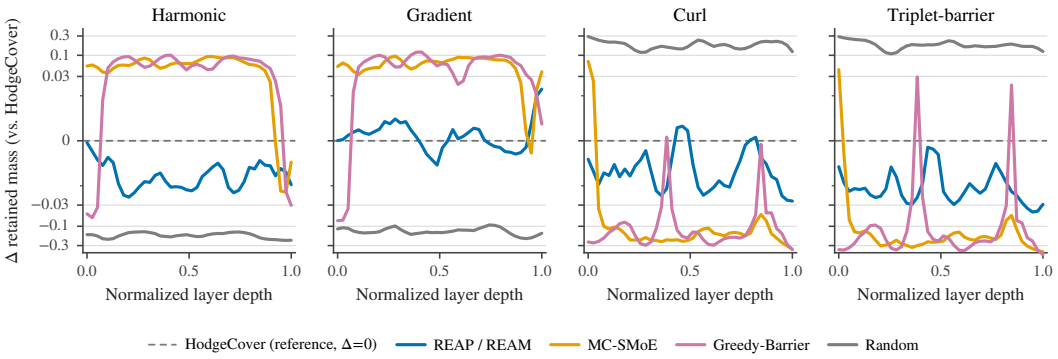

Figure 8: Per-layer deviation from HodgeCover on Qwen 3.5-35B-A3B at 66% across the four Hodge components.

Table 11: Macro retained mass over MoE layers at 66% rate for the four components of the Hodge decomposition: harmonic (H), gradient (G), curl (C), and raw triplet-barrier (T). HodgeCover, REAP, and REAM are positioned across all four; Greedy-Barrier, MC-SMoE, and Random each maximize a strict subset.

Method	OLMoE-1B-7B				Qwen 3.5-35B-A3B				Qwen 3.5-122B-A10B			
	H	G	C	T	H	G	C	T	H	G	C	T
REAP	0.6838	0.6133	0.3859	0.4972	0.7310	0.7009	0.3741	0.5149	0.6831	0.6729	0.4323	0.5716
REAM	0.6838	0.6133	0.3859	0.4972	0.7310	0.7009	0.3741	0.5149	0.6831	0.6729	0.4323	0.5716
Greedy-Barrier	**0.7469**	**0.6535**	0.1917	0.2269	0.7919	0.7572	0.2183	0.3191	0.7552	**0.7353**	0.2268	0.3278
MC-SMoE	0.7202	0.6372	0.2717	0.3273	**0.7960**	**0.7607**	0.2209	0.3232	**0.7641**	0.7347	0.2397	0.3363
Random	0.5761	0.5690	**0.5858**	**0.7382**	0.5683	0.5642	**0.5649**	**0.7138**	0.5692	0.5654	**0.5650**	**0.7143**
HodgeCover (ours)	0.6927	0.6128	0.3781	0.4914	0.7395	0.6998	0.3819	0.5311	0.6910	0.6734	0.4363	0.5774

F.7 Full ablation breakdown

The wrapped Table 2 in the main body reports the Qwen 3.5-35B-A3B ablations. Table 12 below adds the per-task breakdown for all three backbones at both rates.

The cross-backbone pattern matches the Qwen 3.5-35B-A3B reading. On Qwen 3.5-122B-A10B at 66% HodgeCover gains +11.4 pp DS-Avg over Hodge No-Triangle, +8.9 pp over Greedy-Barrier, and +32.2 pp over Triplet-Hypergraph. Across the two Qwen scales the non-Hodge ablations cluster: Hodge No-Triangle is 3–12 pp DS-Avg below HodgeCover and Triplet-Hypergraph collapses by 8–32 pp; on Qwen 3.5-35B-A3B, Triplet-Penalty lands within 1 pp of Greedy-Barrier but 4–6 pp below HodgeCover. On the two Qwen scales the HodgeCover pipeline leads every non-Hodge alternative, the triangle term is necessary for it to deliver, and the soft form of the topological coverage is necessary for the triangle term not to over-veto. On OLMoE-1B-7B, HodgeCover and Triplet-Penalty are within 1 pp DS-Avg at 33%, and at 66% all expert-count rows fall to the 33–36% DS-Avg band.

F.8 Vision-language study

The construction consumes only routing statistics and expert outputs, so it can be applied to a MoE whose routing is conditioned on images. We applied it to MolmoE-1B (Deitke et al., 2024), an open multimodal MoE built on the OLMoE backbone. Calibration uses 512 COCO-2017 image-text pairs (Lin et al., 2014), excluding every evaluation image ID. VQA accuracy is measured with the official VQA accuracy metric on a deterministic 1,000-question subset of the VQAv2 validation set (Goyal et al., 2017); caption negative log-likelihood (NLL) is measured on a disjoint 500-image set.

Vision-conditioned routing carries harmonic structure: the harmonic energy fraction ranges from 24% to 85% across layers, with mean 54%. Table 13 compares HodgeCover and REAP at 33% expert reduction.

HodgeCover is ahead of REAP on VQA accuracy by 3.80 pp, with a paired confidence interval of $[1.96, 5.61]$. On caption NLL the uncompressed model, HodgeCover, and REAP are statistically indistinguishable. Under strict short-answer extraction with greedy decoding, HodgeCover retains

Table 12: Per-task ablation breakdown across all three backbones at 33% and 66% rate, with the rightmost column reporting the unweighted nine-task DS-Avg. Bold = best within each displayed (model, rate) block. Triplet-Penalty is reported on OLMoE and Qwen 3.5-35B-A3B. The rows match Table 2 on Qwen 3.5-35B-A3B in the main body.

Variant	Wiki	C4	ARC-c	ARC-e	BoolQ	HellaS	MMLU	PIQA	TQA	WinoG	GSM8K	DS-Avg
OLMoE-1B-7B — 33% rate												
HodgeCover (ours)	23.50	22.91	**32.8**	50.3	62.8	**67.0**	25.6	73.0	**41.7**	**64.9**	2.2	46.7
Hodge No-Triangle	38.37	37.36	27.6	49.9	62.0	47.2	26.6	66.5	36.1	55.5	2.3	41.5
Triplet-Hypergraph	42.90	41.42	26.6	40.2	61.9	46.8	25.1	69.0	40.9	53.8	2.3	40.7
Triplet-Penalty	**21.03**	**20.17**	32.5	**50.5**	**64.4**	65.2	**33.8**	**73.8**	38.6	63.2	**3.3**	**47.3**
Greedy-Barrier	21.37	20.42	31.6	49.9	62.5	65.7	32.4	73.8	37.7	61.4	3.0	46.4
OLMoE-1B-7B — 66% rate												
HodgeCover (ours)	1,134.9	535.5	24.2	29.5	38.6	30.0	25.4	52.7	47.7	**51.2**	**0.1**	33.3
Hodge No-Triangle	20,714.5	4,596.3	24.5	30.4	40.0	27.8	25.2	51.7	49.9	50.3	0.0	33.3
Triplet-Hypergraph	19,030.7	14,481.7	24.7	27.8	39.9	26.8	25.5	51.9	**50.6**	49.8	0.0	33.0
Triplet-Penalty	**427.3**	**252.7**	**25.2**	30.6	**52.6**	**34.2**	**26.6**	56.1	50.0	50.7	0.0	**36.2**
Greedy-Barrier	435.2	291.8	24.6	**31.5**	48.8	32.4	24.6	**56.4**	49.9	49.6	0.1	35.3
Qwen 3.5-35B-A3B — 33% rate												
HodgeCover (ours)	**9.97**	**13.75**	**56.3**	**73.9**	**91.3**	**82.5**	**80.7**	**82.9**	**54.7**	**76.3**	85.1	**75.9**
Hodge No-Triangle	12.48	17.10	46.4	67.7	87.2	73.3	66.4	77.9	46.8	70.2	73.7	67.7
Triplet-Hypergraph	14.29	16.76	36.3	57.7	88.1	72.4	53.1	78.2	48.8	73.2	72.5	64.5
Triplet-Penalty	10.22	13.82	47.2	66.3	87.0	80.6	74.3	82.0	50.5	74.8	88.0	72.3
Greedy-Barrier	10.22	13.79	46.6	66.3	89.8	80.7	74.1	81.8	50.9	75.7	**88.2**	72.7
Qwen 3.5-35B-A3B — 66% rate												
HodgeCover (ours)	**15.13**	18.86	**46.4**	**67.5**	**88.1**	**72.5**	**54.5**	**78.1**	**51.7**	73.7	**67.5**	**66.7**
Hodge No-Triangle	21.34	28.34	35.8	54.7	78.5	57.4	44.1	70.3	45.4	67.6	42.8	55.2
Triplet-Hypergraph	98.08	154.9	20.8	32.7	62.2	30.7	26.0	56.3	44.4	50.9	1.7	36.2
Triplet-Penalty	16.14	18.77	36.0	53.7	86.6	69.6	44.2	76.1	47.1	**74.0**	61.2	60.9
Greedy-Barrier	15.30	**18.13**	33.4	52.6	86.8	69.9	43.9	77.3	47.6	72.5	62.3	60.7
Qwen 3.5-122B-A10B — 33% rate												
HodgeCover (ours)	**7.21**	**12.81**	**61.7**	**80.1**	74.4	**84.9**	**83.9**	**83.1**	51.6	77.3	**84.5**	**75.7**
Hodge No-Triangle	9.51	15.34	56.3	78.9	72.5	78.1	76.2	80.5	50.2	75.6	82.9	72.4
Triplet-Hypergraph	11.54	15.37	45.1	67.7	**82.5**	76.9	60.1	80.2	49.1	75.1	69.5	67.4
Greedy-Barrier	7.36	12.88	58.6	78.2	73.6	84.7	79.3	82.6	**52.3**	**78.5**	72.1	73.3
Qwen 3.5-122B-A10B — 66% rate												
HodgeCover (ours)	**12.46**	17.21	**52.2**	**71.1**	75.3	**77.0**	**69.0**	**80.1**	**51.5**	**76.4**	**78.1**	**70.1**
Hodge No-Triangle	16.79	22.95	39.9	61.4	**85.7**	66.0	48.9	75.0	43.0	70.6	37.3	58.7
Triplet-Hypergraph	59.87	94.01	23.6	39.2	62.2	33.1	25.5	58.1	43.4	53.0	2.7	37.9
Greedy-Barrier	12.76	**16.29**	38.6	60.8	79.0	75.0	42.2	78.5	47.5	75.4	53.3	61.2

Table 13: MolmoE-1B at 33% expert reduction.

Method	VQA accuracy (%)	Caption NLL
Uncompressed	18.5	4.756
HodgeCover	16.2	4.752
REAP	12.4	4.768

16.2 of the uncompressed 18.5 VQA accuracy against 12.4 for REAP. The study covers one backbone at one rate.

F.9 SYSTEMS CHARACTERIZATION

Table 14 reports the first-run plan-time of all six methods on OLMoE-1B-7B and Qwen 3.5-35B-A3B. Table 15 reports inference throughput at the 66% rate for REAP+Wanda, REAM+Wanda, and HodgeCover+Wanda (and, for reference, HodgeCover without Stage 2 Wanda) on all three backbones.

Table 14: First-run plan-time at the 33% rate, timed as the full planning pipeline of each method under one uniform protocol on the same hardware. Bold = fastest per column. HodgeCover-only (without Stage 2 Wanda) is included for reference.

Method	OLMoE-1B-7B (s)	Qwen 3.5-35B-A3B (s)
REAP+W	4.0	**6.3**
REAM+W	14.0	24.4
STUN+W	**3.9**	6.5
MC-SMoE	34.5	55.8
HodgeCover+W (ours)	12.6	36.6
HodgeCover (ours)	11.2	25.0

Table 15: Post-compression inference throughput (tok/s) at the 66% rate. Bold = highest per column. HodgeCover-only (without Stage 2 Wanda) is included for reference.

Method	OLMoE-1B-7B	Qwen 3.5-35B-A3B	Qwen 3.5-122B-A10B
REAP+W	6,137	5,384	3,616
REAM+W	6,387	**5,571**	**4,119**
HodgeCover+W (ours)	6,506	5,348	3,801
HodgeCover (ours)	**6,615**	5,318	3,817

Plan-time. Table 14 reports plan-time at 33%, the first rate run on each backbone, where HodgeCover builds the simplicial complex, the pairwise and triplet barrier sweeps, and the Hodge decomposition from scratch. Subsequent rates reuse the cached complex; the per-rate survivor selection and redirect run in single-digit seconds. On OLMoE-1B-7B and Qwen 3.5-35B-A3B the timed region is the full first-run planning pipeline of every method, from calibration loading to the finished compression plan, under one protocol on the same hardware. HodgeCover and HodgeCover+Wanda fall within the range spanned by the learning-free baselines at both the 64- and the 256-expert scale. The asymptotics and the timed implementation are described in App. C.5.

Throughput. Post-compression throughput of HodgeCover+Wanda at 66% exceeds REAP+Wanda's on OLMoE and Qwen 3.5-122B and is within 6% on every backbone (Table 15); the largest difference, about 370 tok/s (6.0% relative) on OLMoE-1B-7B, is in HodgeCover+Wanda's favor.