FEATURE EXTRACTION ADVANTAGE OF GRAPH TRANSFORMERS: PRIMAL-DUAL DECOMPOSITION AND RECONSTRUCTION-RISK ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph Transformers (GTs) replace fixed topology-driven aggregation with feature-adaptive attention, yet the source of their feature-extraction advantage remains unclear. In particular, how this adaptivity produces discriminative representations and measurable reconstruction benefits is not understood. This paper develops a unified framework connecting primal–dual decomposition, unrolled sparse coding, and reconstruction-risk analysis. We first prove that linearized GT attention admits an exact finite decomposition: key-induced primal atoms analyze the input, while query-induced dual atoms synthesize the output from the resulting projection coefficients. We extend this viewpoint to masked softmax attention through an exact normalized exponential-kernel expansion. We then show that removing feature dependence from the query–key coordinates reduces attention to a fixed linear graph filter, with GCN and spectral graph convolution arising as topology-driven special cases. This identifies the boundary between adaptive attention and fixed aggregation. Motivated by the decomposition, we interpret a deep GT as an unrolled proximal-gradient solver for an $\ell_{2,1}$-regularized coding problem: attention generates an adaptive dictionary, and successive layers perform residual correction and group-sparse coefficient refinement. Finally, using a ridge surrogate, we derive explicit lower bounds on the reconstruction-risk advantage of adaptive GT-inspired filters over fixed GCN filters. The shared-basis specialization isolates spectral adaptivity by allowing the two filter families to differ only through their spectral weights. Controlled experiments verify the decompositions and the degeneration to fixed filtering, and show that adaptive attention improves class discriminability and lowers reconstruction risk relative to fixed topology-driven aggregation.

1 INTRODUCTION

Graph-structured data arise naturally in citation networks, recommendation systems, molecular modeling, knowledge graphs, and a wide range of other applications. Graph neural networks (GNNs) provide a general framework for learning from such data by repeatedly propagating and aggregating information over graph neighborhoods, and graph convolutional networks (GCNs) Kipf & Welling (2017) have become a fundamental architecture owing to their simplicity and strong empirical performance. A typical GCN layer propagates features using a matrix that depends only on the graph structure, and then applies a learned transformation. All input signals follow the same aggregation pattern. This topology-driven inductive bias is effective when edges faithfully reflect feature or label similarity, but restrictive under noisy edges, heterophily, heterogeneous local structures, or neighborhoods with only partially informative nodes He et al. (2024); Yao et al. (2022); Li et al. (2024c). Repeated application of a fixed filter can also cause excessive feature mixing and over-smoothing Oono & Suzuki (2020); Zhou et al. (2021); Yang et al. (2024). Graph transformers (GTs) Vaswani et al. (2017); Veličković et al. (2018) overcome this limitation by replacing the fixed propagation operator with feature-dependent attention: query–key comparisons yield an input-specific aggregation matrix that highlights informative interactions while suppressing misleading ones, ensuring that the resulting filter is determined by the informative node features rather than solely by the graph topology. Such adaptivity has led to strong performance in node

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.

and graph classification, heterogeneous graph learning, and large-scale graph representation learning Ying et al. (2021); Kim et al. (2022); Wu et al. (2023); Park et al. (2025); Liang et al. (2026, DOI: 10.1109/TPAMI.2026.3682858); see also recent surveys Chen et al. (2024); Shehzad et al. (2026, DOI: 10.1109/TNNLS.2025.3646122).

The impressive empirical success of GTs has motivated substantial effort to ground their abilities in theoretical foundations. The existing theory spans several complementary directions. Expressivity has been studied through the Weisfeiler–Leman (WL) isomorphism test, ranging from tokenization schemes as expressive as 2-IGN Kim et al. (2022) to WL-aligned model classes Zhu et al. (2023); Müller & Morris (2024). In parallel, spectral attention has been shown to yield universality Kreuzer et al. (2021), and logic-based characterizations have been developed in Ahvonen et al. (2026). The generalization and stability of positional encodings have been analyzed Li et al. (2024a); Huang et al. (2024); depth–width tradeoffs on graph tasks have been characterized Yehudai et al. (2026); and architecture-level connections to graph convolution have been established Ying et al. (2021); Wu et al. (2023). Within the broader transformer literature, attention has also been related to classical neural networks, from convolutional neural networks Cordonnier et al. (2020) to recurrent neural networks Jiang & Li (2024). Despite these advances, we still lack a feature-extraction–level explanation of how a GT layer extracts features, how fixed graph topologies are incorporated into attention, and how to quantify filter adaptivity.

Motivated by this gap, we study the attention itself and ask what feature adaptivity contributes and where it comes from. Answering this question requires more than an expressiveness comparison: it calls for a description of the representation generated by attention, an optimization interpretation, and a quantitative criterion for comparing adaptive and fixed filters. In terms of concreteness, three questions remain open. (i) Is it possible to break down a linearized GT layer into clear analysis and synthesis operators that make the functions of the key, query, and value components explicit? (ii) Does this kind of decomposition also hold for the masked softmax operator, or does it arise only from linearizing the attention weights? (iii) Can it be linked to an optimization framework that clarifies the impact of depth and allows a principled comparison between adaptive GT and fixed GCN filters? These questions are interconnected, and addressing them together provides a complete operator-level description of feature-adaptive attention.

We tackle these questions using an integrated framework that draws on primal–dual decomposition, unrolled group-sparse coding, and reconstruction-risk analysis. We begin by interpreting a GT layer as a data-driven dictionary-coding process, where the input is compared to an adaptive dictionary and then reconstructed via the query, key, and value branches. From this operator-level viewpoint, the following contributions emerge:

- **Attention decomposition.** Exact primal–dual expansions for linearized and masked-softmax attention identify key-induced analysis atoms and query-induced synthesis atoms.
- **Filter-family characterization.** Feature-independent attention reduces to a fixed graph filter, with GCN and spectral convolution as special cases.
- **Optimization interpretation.** GT depth is interpreted as unrolled proximal-gradient refinement for an $\ell_{2,1}$-regularized coding problem, enabling comparison with fixed-filter GCN solvers.
- **Quantitative comparison.** A ridge surrogate yields lower bounds on the reconstruction-risk advantage of adaptive filters, including a shared-basis regime that isolates adaptive spectral weighting. Controlled experiments verify the predicted mechanisms.

2 Related Work

2.1 Theoretical analysis of transformers

Transformer theory first focused on approximation and expressivity, including universal approximation under dense, sparse, low-rank, and constrained attention Yun et al. (2019; 2020); Kajitsuka & Sato (2024); Kratsios et al., and later quantified approximation rates and structural limitations Gurevych et al. (2022); Takakura & Suzuki (2023); Jiang & Li (2024); Wang & E (2024); Bhojanapalli et al. (2020); Dong et al. (2021). A complementary line interprets transformers through classical algorithms and architectures: attention can implement optimization procedures Bai et al.

(2023), approximate convolution Cordonnier et al. (2020); Li et al. (2021), and represent solution maps of deep sparse-coding problems Li et al. (2024b; 2026). These results motivate our optimization view, but they do not directly account for graph topology or feature-adaptive graph filtering.

2.2 Theoretical analysis of graph transformers

Graph transformers incorporate topology through positional or structural encodings, including Graphormer Ying et al. (2021), GraphiT Mialon et al. (2021), EGT Hussain et al. (2022), spectral attention Kreuzer et al. (2021), and SGFormer Wu et al. (2023).

Theory has established generalization and stability properties Li et al. (2024a); Huang et al. (2024), depth–width tradeoffs Yehudai et al. (2026); Zhao et al. (2023), and expressivity through 2-IGN, spectral universality, WL tests, and logic-based characterizations Kim et al. (2022); Kreuzer et al. (2021); Zhu et al. (2023); Zhou et al. (2024); Müller & Morris (2024); Ahvonen et al. (2026). Architectural analyses further show that structural encodings can realize GNN operations Ying et al. (2021), relate attention to graph-signal denoising Wu et al. (2023), or interpret attention as adaptive spectral filtering Wi et al. (2025). These studies explain what GTs can represent and how topology can enter the architecture. They do not, however, decompose standard attention into explicit analysis and synthesis operators or quantify adaptive and fixed filters under a common reconstruction criterion. Our work addresses precisely this operator-level gap.

3 Preliminaries and Problem Setup

Let $G = (\mathcal{V}, \mathcal{E})$ be an undirected graph with $|\mathcal{V}| = N$. The node-feature matrix is $X = [x_1, \dots, x_N]^\top \in \mathbb{R}^{N\times d}$, where $x_i \in \mathbb{R}^d$ is the feature vector of node $i \in \mathcal{V}$, and we assume $\|x_i\|_2 = 1$ for notational simplicity. We consider learning on graph-structured data in general, although the analysis below only requires a representation-learning viewpoint. A single-head transformer layer uses matrices $Q, K : \mathbb{R}^{\ell\times d}$ and $V : \mathbb{R}^{d'\times d}$ to form query, key, and value vectors $q_i = Qx_i$, $k_j = Kx_j$, $u_j = Vx_j$. With full attention, the attention coefficient from node j to node i is

$$\alpha_{ij} = \frac{\exp(\langle q_i, k_j\rangle)}{\sum_{m=1}^{N}\exp(\langle q_i, k_m\rangle)},$$

and the output feature is $x_i' = \sigma\Big(\sum_{j=1}^{N}\alpha_{ij}u_j\Big)$, where σ is an activation function. For simplicity, our analysis omits the feed-forward layers of the transformer proposed in Vaswani et al. (2017). However, our analysis can be directly extended to incorporate full transformer layers.

Furthermore, to isolate the bilinear feature-selection mechanism of attention, our decomposition result is stated for the linearized aggregation operator

$$\bar{x}_i' := \sum_{j=1}^{N}\langle q_i, k_j\rangle u_j = \sum_{j=1}^{N}\langle Qx_i, Kx_j\rangle Vx_j. \tag{1}$$

In contrast to the Graph Transformer, which relies on learnable pairwise attention, graph convolution operates via a fixed, deterministic neighborhood averaging scheme. Let $A \in \mathbb{R}^{N\times N}$ denote the symmetric adjacency matrix with non-negative entries $a_{ij} \geq 0$. A single graph convolution layer is typically defined as $X' = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}XW^\top\right)$, where $\tilde{A} = A + I_N$, $\tilde{D}_{ii} = \sum_j \tilde{a}_{ij}$, and W contains trainable parameters Kipf & Welling (2016). This formulation originates from a first-order approximation of spectral graph convolutions via Chebyshev polynomials. To isolate the core aggregation mechanism from normalization and nonlinearities, we consider the simplified message-passing update for node i: $\bar{x}_i' = \sum_j a_{ij}Wx_j$, or, equivalently, in matrix form $X' = AXW^\top$. For simplicity, our analysis focuses on this simplified variant; nevertheless, our theoretical results are general and readily extend to the renormalized formulation.

Our comparisons use two objectives introduced where needed below: an $\ell_{2,1}$-regularized coding objective for the unrolled architecture and a ridge surrogate for explicit reconstruction-risk bounds.

4 MAIN THEORETICAL RESULTS

We write $\widetilde{X} := \|_{i=1}^{N} x_i$ for the vertical concatenation of all node features and, for the concatenated vectors $A = \|_{i=1}^{N} A_i$ and $B = \|_{i=1}^{N} B_i$ with blocks in $\mathbb{R}^d$, we define $\langle A, B\rangle := \sum_{i=1}^{N}\langle A_i, B_i\rangle$. Let $h_{q,s}(x) := e_s^\top Qx$ and $h_{k,s}(x) := e_s^\top Kx$ denote the s-th coordinates of the query and key projections. Let $v_t \in \mathbb{R}^d$ denote the t-th row of V viewed as a column vector. Proofs of all theoretical results in this section are provided in Appendix A.

The following theorem shows that linearized attention first measures the input against key-induced primal atoms and then synthesizes the output through query-induced dual atoms.

Theorem 1 (Primal–Dual Feature Decomposition). *Define $\widetilde{Y} := \|_{i=1}^{N} \left(V^\top \bar{x}_i'\right) \in \mathbb{R}^{Nd}$, where $\bar{x}_i'$ is given by equation 1. For $s \in \{1, \dots, \ell\}$ and $t \in \{1, \dots, d'\}$, define $\phi_{s,t}(\widetilde{X}) := \|_{j=1}^{N} \left(h_{k,s}(x_j)v_t\right)$, $\psi_{s,t}(\widetilde{X}) := \|_{i=1}^{N} \left(h_{q,s}(x_i)v_t\right)$. Then, we have*

$$\widetilde{Y} = \sum_{s=1}^{\ell}\sum_{t=1}^{d'}\langle \widetilde{X}, \phi_{s,t}(\widetilde{X})\rangle\, \psi_{s,t}(\widetilde{X}). \tag{2}$$

The decomposition also provides a direct interpretation of the reconstruction dictionary used in the coding problem. Specifically, the dual factors $\psi_{s,t}(\tilde{X})$ play the role of data-adaptive dictionary atoms, while the associated code coefficients are given by the primal projections $\langle \tilde{X}, \phi_{s,t}(\tilde{X})\rangle$.

We next show that the same viewpoint extends to the masked softmax attention operator. The main difference is that the softmax exponential lifts the finite bilinear decomposition to an infinite-dimensional exponential-kernel decomposition, while the row-wise normalization appears as a node-dependent scaling factor.

Let $\mathcal{N}_i$ denote the attention neighborhood of node i, and define the masked softmax pre-activation output $\bar{x}_i^{\mathrm{sm}} := \sum_{j\in\mathcal{N}_i} \alpha_{ij} V x_j$, $\alpha_{ij} := \frac{\exp(\langle Qx_i, Kx_j\rangle)}{\mathcal{Z}_i}$, where $\mathcal{Z}_i := \sum_{m\in\mathcal{N}_i}\exp(\langle Qx_i, Kx_m\rangle)$. Parallelly, For a multi-index $\nu = (\nu_1, \dots, \nu_\ell) \in \mathbb{N}_0^\ell$, we define $|\nu| := \sum_{s=1}^{\ell}\nu_s$, $\nu! := \prod_{s=1}^{\ell}\nu_s!$, and $h_q^\nu(x_i) := \prod_{s=1}^{\ell} h_{q,s}(x_i)^{\nu_s}$,, $h_k^\nu(x_j) := \prod_{s=1}^{\ell} h_{k,s}(x_j)^{\nu_s}$. For each $i \in [N]$, $\nu \in \mathbb{N}_0^\ell$, and $t \in \{1, \dots, d'\}$, define $\phi_{i,\nu,t}^{\mathrm{sm}}(\tilde{X}) := \left\|_{j=1}^{N} 1_{\{j\in\mathcal{N}_i\}} h_k^\nu(x_j)v_t$, and $\psi_{i,\nu,t}^{\mathrm{sm}}(\tilde{X}) := \left\|_{r=1}^{N} 1_{\{i=r\}} \frac{h_q^\nu(x_i)}{\nu!\mathcal{Z}_i} v_t$, where $1_A = 1$ if event A holds, $= 0$ otherwise.

Theorem 2 (Primal–dual decomposition of masked softmax attention). *Let $\tilde{Y}^{\mathrm{sm}} := \left\|_{i=1}^{N} V^\top \bar{x}_i^{\mathrm{sm}}$ be the concatenation of pre-activation output features. Then the masked softmax attention output admits the expansion*

$$\tilde{Y}^{\mathrm{sm}} = \sum_{i=1}^{N}\sum_{\nu\in\mathbb{N}_0^\ell}\sum_{t=1}^{d'}\left\langle \tilde{X}, \phi_{i,\nu,t}^{\mathrm{sm}}(\tilde{X})\right\rangle \psi_{i,\nu,t}^{\mathrm{sm}}(\tilde{X}).$$

Unlike graph-filtering approaches that redesign attention in the spectral domain Wi et al. (2025), Theorems 1–2 characterize the standard GT attention operator itself through exact primal–dual decompositions. To further enable node-specific attention weights, we augment each feature vector with a one-hot positional encoding: $z_i = [x_i^\top, e_i^\top]^\top \in \mathbb{R}^{d+N}$, where $e_i \in \mathbb{R}^N$ is the i-th standard basis vector. Let $Z = [z_1, \dots, z_N]^\top \in \mathbb{R}^{N\times(d+N)}$ denote the augmented input matrix.

Corollary 1 (Degeneracy to a fixed graph filter; GCN as a special case). *Let $A \in \mathbb{R}^{N\times N}$ be any target matrix (e.g., the normalized adjacency matrix of a given graph), and let $W \in \mathbb{R}^{d'\times d}$ be any desired value projection. Then there exist matrices $Q, K \in \mathbb{R}^{\ell\times(d+N)}$, $V \in \mathbb{R}^{d'\times(d+N)}$ such that the linearized, single-head Graph Transformer update equation 1 on the augmented features reduces exactly to $\bar{X}' = AXW^\top$, where $\bar{X}' = [\bar{x}_1', \dots, \bar{x}_N']^\top \in \mathbb{R}^{N\times d'}$ is the output using only the original node features.*

The corollary identifies the source of the additional flexibility of GTs: When query and key representations are restricted to depend only on positional information, the GT layer loses feature adaptivity and reduces to a fixed graph filter. Hence, GCN can be viewed as a restricted subfamily of GT. In

particular, if A is chosen as the normalized adjacency matrix $\hat{A}$, or a graph-polynomial/spectral filter $g(L)$, the update recovers a GCN-like graph convolution. Graphormer Ying et al. (2021) shows that suitable structural encodings allow transformers to represent several GNN operations. Corollary 1 instead gives an explicit operator-level construction showing that feature-independent query–key coordinates recover an arbitrary fixed linear graph filter.

Next, based on the decomposition, we show that a GT layer implements one learned proximal-gradient step for the current attention-induced dictionary. A deep GT is therefore an unrolled adaptive solver, analogous to learned sparse-coding networks.

Theorem 3 (Deep GT as an unrolled adaptive group-sparse solver)**.** *For a graph filter $D_\theta \in \mathbb{R}^{N\times N}$ define $\mathcal{J}(Z) := \frac{1}{2}\|XV^\top - D_\theta Z\|_F^2 + \alpha\|Z\|_{2,1}$. For a sequence of filters $\{D_\theta^k\}_{k=0}^{K-1}$, define the recurrence*

$$\widetilde{Z}^{k+1} = Z^k + \eta_k (D_\theta^k)^\top (XV^\top - D_\theta^k Z^k),$$
$$Z_{:,j}^{k+1} = \left(1 - \frac{\eta_k \alpha}{\|\widetilde{Z}_{:,j}^{k+1}\|_2}\right)_+ \widetilde{Z}_{:,j}^{k+1}, \qquad j = 1, \dots, d'.$$

1. *If $D_\theta^k = D_\theta$ is fixed for all k, this recurrence is exactly the proximal-gradient method for minimizing $\mathcal{J}(Z)$.*
2. *If D_θ^k is generated by query–key attention, then a deep GT block implements an unrolled adaptive group-sparse coding solver.*
3. *If D_θ^k is restricted to a fixed topology-driven graph filter, the same recurrence reduces to a GCN-type unrolled sparse-coding solver.*

Whereas SGFormer Wu et al. (2023) interprets a single attention layer as an optimization step for graph-signal denoising, Theorem 3 interprets depth as repeated proximal-gradient refinement of an $\ell_{2,1}$-regularized coding objective with attention-induced filters. As a consequence, a clear and general performance comparison between GT and GCN is straightforward.

Corollary 2 (Induced-loss comparison between GT and GCN)**.** *Let $\mathcal{Z}_{\mathrm{GCN}}^{(K)}$ and $\mathcal{Z}_{\mathrm{GT}}^{(K)}$ denote the sets of representations reachable after K unrolled $\ell_{2,1}$-proximal steps using filters from the GCN and GT families, respectively. Then $\mathcal{Z}_{\mathrm{GCN}}^{(K)} \subseteq \mathcal{Z}_{\mathrm{GT}}^{(K)}$. Consequently, for any task loss ℓ, $\inf_{Z\in\mathcal{Z}_{\mathrm{GT}}^{(K)}} \ell(Z) \le \inf_{Z\in\mathcal{Z}_{\mathrm{GCN}}^{(K)}} \ell(Z)$.*

To compare fixed and adaptive filter families explicitly, we use a ridge reconstruction surrogate. It does not model the complete layer-wise dynamics; rather, it isolates the effect of enlarging the admissible filter family from topology-driven GCN filters to feature-adaptive GT filters. Specifically, for a graph filter D_θ, define

$$Z(\theta) := \arg\min_{Z\in\mathbb{R}^{N\times d}} \|X - D_\theta Z\|_F^2 + \alpha\|Z\|_F^2, \tag{3}$$

where we absorb $V^\top$ into X and set $\mathcal{R}(Z) := \|Z\|_F^2$. For the data model $X = X^* + \eta$ with a target X^* and noise η, we evaluate the risk of reconstruction by

$$\ell(Z(\theta)) := \|X^* - D_\theta Z(\theta)\|_F^2, \tag{4}$$

which encourages the filtered features to remain close to the target feature matrix X^*. When the graph convolution is given by $D_\theta := A$, we denote the minimizer of equation 3 by Z_A. When, instead, the graph filter is induced by the graph transformer with positional encoding as in Corollary 1, we denote the corresponding minimizer by Z_{GT}.

The following theorem therefore quantifies the reconstruction-risk advantage of the adaptive filter family, rather than the layer-wise optimization trajectory of a specific GT architecture.

Theorem 4 (Ridge Surrogate Comparison of Reconstruction-Risk)**.** *Let $A = H\cdot\mathrm{Diag}(\lambda_1, \dots, \lambda_N)\cdot H^\top$ be a fixed graph convolution matrix where H is orthonormal and $|\lambda_1| \ge \cdots \ge |\lambda_N|$. We work with the data model $X = X^* + \eta$, where $X^* \in \mathbb{R}^{N\times d}$ contains target features and $\eta \in \mathbb{R}^{N\times d}$ is random noise with $\mathbb{E}[\eta] = \mathbf{0}$ and $\mathbb{E}[\eta\eta^\top] = \sigma^2 d\, I_N$.*

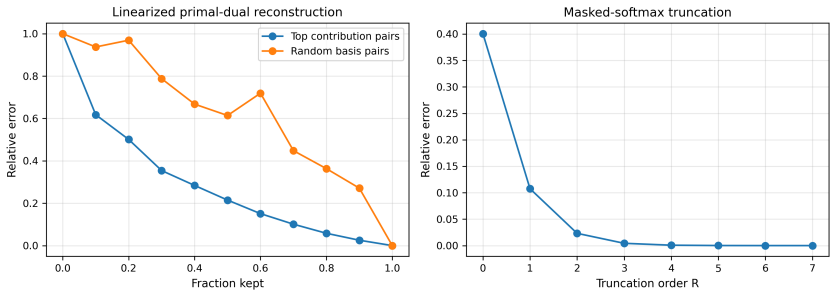


Figure 1: Experiment 1: structural verification of the primal–dual decompositions. Left: relative reconstruction error for the linearized GT output when retaining an increasing fraction of primal–dual basis pairs. Top-ranked pairs reconstruct the output more efficiently than randomly selected pairs, and the error vanishes when all pairs are retained. Right: relative approximation error of the finite-order masked-softmax expansion. The error decreases rapidly with the truncation order, consistent with Theorem 2.

Denote the reconstruction gap by $\Delta = \mathbb{E}[\ell(Z_A)] - \mathbb{E}[\ell(Z_{GT})]$. *Then there exist a GT filter such that the reconstruction gap achieves*

$$\Delta \geq \sup_{P:P^\top P=I_N} \sum_{i=1}^{N} \left[\frac{\alpha^2 \|H_{:,i}^\top X^*\|_2^2 + \sigma^2 d\lambda_i^4}{(\lambda_i^2+\alpha)^2} - \frac{\sigma^2 d\|P_{:,i}^\top X^*\|_2^2}{\sigma^2 d + \|P_{:,i}^\top X^*\|_2^2} \right].$$

Theorem 4 characterizes a lower bound on the reconstruction gap between the GT and GCN filters. As also indicated by Theorem 1 and Corollary 1, we can see the advantage that the GT filter is capable of learning any primal-dual decomposition, encompassing all orthonormal bases. Performing optimization over all orthonormal bases leads to a tighter lower bound.

The following corollary provides a more concise lower bound in the case where the GT filter recovers the orthonormal system determined by the GCN filter.

Corollary 3. *Suppose that the conditions of Theorem 4 are satisfied. The lower bound in Theorem 4 attains its maximum when the columns of* P *are eigenvalues of* $X^*(X^*)^\top$. *More precisely, the reconstruction gap* Δ *obeys the following lower bound*

$$\Delta \geq \sum_{i=1}^{N} \frac{\alpha^2 \|H_{:,i}^\top X^*\|_2^2 + \sigma^2 d\lambda_i^4}{(\lambda_i^2+\alpha)^2} - \sum_{i=1}^{N} \frac{\sigma^2 d\mu_i}{\sigma^2 d + \mu_i} \geq \frac{(\alpha\|AX^*\|_F^2 - \sigma^2 d\|A^2\|_F^2)^2}{(\lambda_1^2+\alpha)^2(\|X^*\|_F^2 + \sigma^2 d)\|A^2\|_F^2}, \tag{5}$$

where $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_N \geq 0$ *are the eigenvalues of* $X^*(X^*)^\top$ *and the last expression is obtained when* P *is formed from the eigenvectors of* A.

Theorems 4 and Corollary 3 provide a family of lower bounds on $\Delta = \mathbb{E}[\ell(Z_A)] - \mathbb{E}[\ell(Z_{\text{GT}})]$. Unlike SGFormer Wu et al. (2023), which characterizes attention through a graph-signal-denoising objective, Theorem 4 compares fixed and adaptive filter families under the same ridge surrogate and gives an explicit lower bound on their reconstruction-risk gap.

The two sums in equation 5 compare the energy of X^* in the eigenbasis of the fixed filter with a signal-dependent baseline obtained from the eigenbasis of $X^*(X^*)^\top$. Thus, their difference quantifies a basis-misalignment penalty. In the noiseless case, directions with both large $|\lambda_i|$ and large signal projection dominate the bound. The simplified bound is strictly positive when $\alpha\|AX^*\|_F^2 > \sigma^2 d\|A^2\|_F^2$. The guarantee is one-sided and depends on the unknown clean signal X^*; a zero lower bound does not imply a zero reconstruction gap, and reconstruction risk need not coincide with downstream-task performance. Further interpretation, including the noiseless specialization, is deferred to Appendix B.

5 Experimental Study

We conduct controlled numerical experiments to test the mechanisms developed above rather than pursue state-of-the-art benchmark performance. The experiments address three questions. First, do

Table 1: Experiment 1: representative structural approximation errors. The top-ranked primal–dual pairs reconstruct the linearized output substantially more efficiently than randomly selected pairs. For masked softmax, the finite-order truncation error decreases rapidly as the order R increases.

Linearized primal–dual reconstruction		
Fraction kept	Top pairs	Random pairs
0.2	0.501	0.969
0.4	0.284	0.668
0.6	0.151	0.719
0.8	0.058	0.363
1.0	0.000	0.000

Masked-softmax finite-order approximation	
Order R	Relative error
0	0.4007
1	0.1075
2	0.0232
4	0.000735
7	0.000002

linearized and masked-softmax attention exhibit the primal–dual structure predicted by Theorems 1 and 2? Second, does gradually removing feature dependence continuously reduce attention to the fixed-filter regime of Corollary 1? Third, does enlarging the adaptive filter family reduce ridge reconstruction risk as predicted by Theorem 4 and Corollary 3? Additional details are provided in Appendix C.

5.1 Synthetic Data and Evaluation Protocol

Each graph contains $N = 120$ nodes with balanced labels $y_i \in \{-1, +1\}$ and feature dimension $d = 16$. A fraction $\gamma \in \{0.1, 0.2, 0.4, 0.6, 0.8\}$ of nodes carries a strong label-aligned component; the remaining nodes contain a weak label signal, nuisance features, and Gaussian perturbations. Thus, γ controls the richness of discriminative information. Edges follow a noisy, weakly homophilous topology with additional random edges, making topology-only aggregation imperfect and allowing us to test whether attention selects informative feature directions.

All results are averaged over three random seeds and reported as mean $\pm$ standard deviation. Compared settings use matched graph realizations and train/test splits, with a 20% label rate. A ridge classifier evaluates the learned node representations. We also report the Fisher ratio $\mathcal{F}(Z) = \frac{\mathrm{Tr}(S_b)}{\mathrm{Tr}(S_w)+\epsilon}$, where S_b and S_w are the between-class and within-class scatter matrices. Larger $\mathcal{F}(Z)$ indicates stronger class separation. Auxiliary diagnostics, including pairwise cosine similarity and graph-filter distance, provide mechanistic rather than independent performance measures.

5.2 Experiment 1: Primal–Dual and Masked-Softmax Decomposition

For a GT layer with projections Q, K, V, let $Q_X = XQ^\top$, $K_X = XK^\top$, and $V_X = XV^\top$. The linearized pre-activation output is $H_{\mathrm{lin}} = (Q_X K_X^\top)V_X = Q_X C$, $C = K_X^\top V_X$. Each entry $C_{s,t}$ determines one primal–dual contribution. We rank these contributions by norm and compare reconstruction from the top-ranked pairs against uniformly selected pairs with the same budget. For m retained pairs, the relative error is

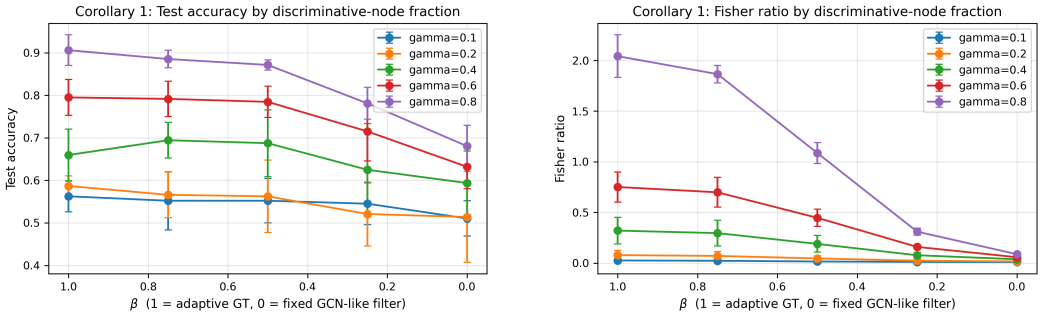
$$\mathrm{Err}_{\mathrm{lin}}(m) = \frac{\|H_{\mathrm{lin}} - H_{\mathrm{lin}}^{(m)}\|_F}{\|H_{\mathrm{lin}}\|_F}.$$

For masked softmax, we replace $\exp(s_{ij})$ by $\sum_{r=0}^{R} s_{ij}^r / r!$ inside the masked-softmax numerator and compare the resulting truncated output with the exact masked-softmax output.

Figure 1 and Table 1 confirm both predictions. Linearized reconstruction is exact when all pairs are retained, directly verifying the finite identity in Theorem 1. The contribution ranking is also informative: retaining the top 20% gives error 0.501 rather than 0.969 for random selection, and the top-80% error is only 0.058. For masked softmax, the approximation error falls from 0.4007 at $R = 0$ to 0.0232 at $R = 2$ and approximately 2×10^{-6} at $R = 7$. The rapid decay supports the finite-order approximation implied by the exact expansion in Theorem 2.

5.3 Experiment 2: Degeneration from GT to GCN-like Filtering

Corollary 1 predicts that removing feature dependence from the query–key coordinates reduces a GT layer to a fixed topology-driven filter. We test this transition continuously with $D_\beta = \beta D_{\mathrm{attn}} +$



(a) Test accuracy. (b) Fisher ratio.

Figure 2: Experiment 2 across discriminative-node fractions γ. Moving from $\beta = 1$ to $\beta = 0$ removes feature adaptivity and generally reduces accuracy and class separation, with the largest effect when discriminative information is abundant.

$(1-\beta)A$, $\beta \in [0,1]$, followed by row normalization and two filtering steps. Here D_{attn} is feature-adaptive, A is the row-normalized adjacency, $\beta = 1$ is the adaptive endpoint, and $\beta = 0$ is the fixed GCN-like endpoint. The direct diagnostic is

$$\text{Dist}_{\text{fixed}}(\beta) = \frac{\|D_\beta - A\|_F}{\|A\|_F},$$

which should vanish at $\beta = 0$. Accuracy and Fisher ratio measure the downstream cost of removing feature adaptivity.

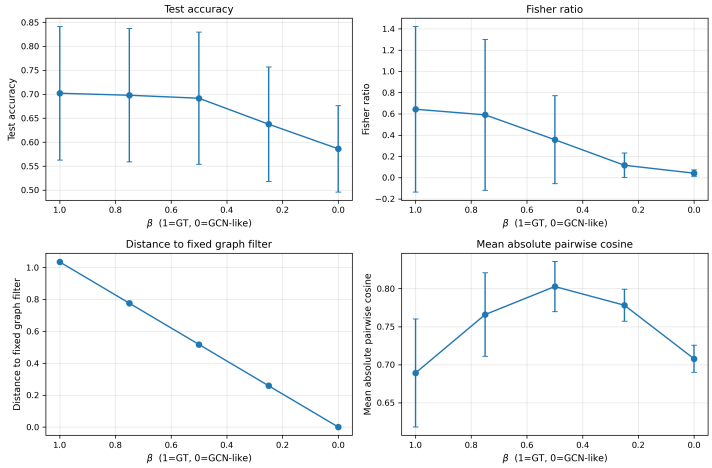


Figure 3: Experiment 2: diagnostics along the degeneration path. The filter distance reaches zero at $\beta = 0$, while accuracy and Fisher ratio decrease as feature adaptivity is removed.

Table 2: Experiment 2 results ($\gamma = 0.8$).

β	Accuracy	Fisher ratio	$\text{Dist}_{\text{fixed}}$	Pairwise cosine
1.00	0.906 ± 0.036	2.042 ± 0.211	1.036 ± 0.003	0.584 ± 0.022
0.75	0.885 ± 0.021	1.866 ± 0.086	0.777 ± 0.003	0.686 ± 0.022
0.50	0.872 ± 0.012	1.087 ± 0.104	0.518 ± 0.002	0.761 ± 0.024
0.25	0.781 ± 0.038	0.310 ± 0.034	0.259 ± 0.001	0.770 ± 0.023
0.00	0.681 ± 0.049	0.087 ± 0.008	0.000 ± 0.000	0.708 ± 0.022

Table 3: Experiment 3 risk gaps.

γ	Filter-set size ratio $\lvert\mathcal{A}\rvert/\lvert\mathcal{B}\rvert$			
	1.0	2.0	3.0	4.0
0.1	0.0 ± 0.0	2.9 ± 1.4	6.5 ± 3.0	8.0 ± 3.6
0.4	0.0 ± 0.0	9.9 ± 0.9	26.7 ± 7.0	33.2 ± 9.4
0.8	0.0 ± 0.0	23.3 ± 2.5	60.5 ± 8.9	71.1 ± 9.9

Figures 2–3 and Table 2 verify both the algebraic endpoint and its practical cost. The distance to A decreases to zero at $\beta = 0$, exactly realizing the fixed-filter limit. At $\gamma = 0.8$, accuracy falls

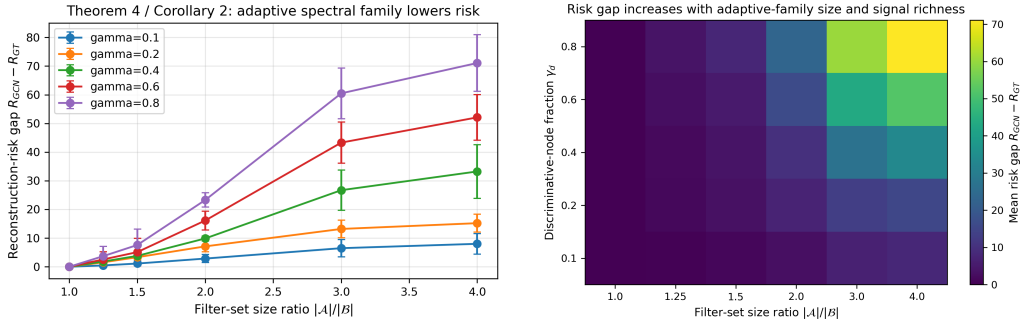
from 0.906 to 0.681 and the Fisher ratio from 2.042 to 0.087. The decline is clearest for larger γ, where attention has more discriminative information to exploit. Pairwise cosine similarity is non-monotone, indicating that the loss of class separation is not captured by a single conventional oversmoothing statistic.

5.4 Experiment 3: Reconstruction-Risk Advantage of Adaptive Spectral Families

For noisy input $X = X^* + \eta$, we evaluate the ridge code $Z_D = (D^\top D + \alpha I)^{-1} D^\top X$ using $\ell(D) = \|X^* - DZ_D\|_F^2$. We compare the best fixed-family and adaptive-family risks:

$$R_{\mathrm{GCN}} = \min_{D \in \mathcal{B}} \ell(D), \qquad R_{\mathrm{GT}} = \min_{D \in \mathcal{A}} \ell(D). \tag{6}$$

The plotted gap is $\Delta R = R_{\mathrm{GCN}} - R_{\mathrm{GT}}$. The fixed family is $\mathcal{B} = \{A, A^2, A^3, A^4\}$. Writing $A = H\,\mathrm{Diag}(\lambda)H^\top$, we enlarge it with adaptive filters $\widetilde{A}_m := H\,\mathrm{Diag}(\theta^{(m)})H^\top$ and hence obtain $\mathcal{A} := \mathcal{B} \cup \{\widetilde{A}_m\}_{m=1}^M$. All candidates therefore share the same eigenvectors and differ only in spectral weights, isolating adaptivity from a change of basis.



(a) Reconstruction-risk gap. (b) Gap heatmap.

Figure 4: Experiment 3: reconstruction-risk advantage of adaptive spectral families. The risk gap $R_{\mathrm{GCN}} - R_{\mathrm{GT}}$ is zero when the adaptive and fixed families coincide, and becomes positive as shared-basis adaptive filters are added. The gap is larger when γ is larger, indicating that adaptive spectral weighting is most useful when more discriminative feature energy is present.

Figure 4 and Table 3 support Theorem 4 and Corollary 3. The gap is zero when the two families coincide and becomes positive as shared-basis adaptive filters are added. At $\gamma = 0.4$, it increases from 0.0 at ratio 1.0 to 9.9, 26.7, and 33.2 at ratios 2.0, 3.0, and 4.0. At ratio 4.0, the gap grows from 8.0 for $\gamma = 0.1$ to 71.1 for $\gamma = 0.8$, showing that adaptive weighting is most useful when more discriminative signal is available.

The component curves in Appendix 5 identify the source of the gap. Since $\mathcal{B}$ does not change, R_{GCN} remains constant as $|\mathcal{A}|/|\mathcal{B}|$ increases. In contrast, R_{GT} decreases when additional adaptive filters become available. Because all candidates retain the same eigenbasis, the improvement cannot be attributed to changing eigenvectors, the graph, or the target matrix; it isolates the value of a richer set of spectral weights.

6 Conclusion

We developed a unified account of the feature-extraction advantage of Graph Transformers. Exact primal–dual decompositions expose the analysis and synthesis roles of attention; feature-independent coordinates recover fixed GCN and spectral filters; and an unrolled $\ell_{2,1}$ formulation gives network depth an optimization interpretation. A ridge surrogate then quantifies the reconstruction-risk cost of restricting aggregation to a fixed filter family. Controlled experiments verify each mechanism and show that adaptive spectral weighting can help even with a shared eigenbasis. Extending the risk analysis to the full group-sparse objective is an important next step.

7 AI USE STATEMENT

During the preparation of this work, the authors used ChatGPT-5.5 and DeepSeek in order to 1) refine the academic language and improve the readability of the manuscript text, and 2) generate initial code skeletons for synthetic experiments and assist in debugging runtime errors. The authors manually verified the logic and output of all AI-generated code, and reviewed all AI-suggested textual revisions for accuracy and scientific validity. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the final content, including all code and experimental data, of the published article.

REFERENCES

Veeti Ahvonen, Maurice Funk, Damian Heiman, Antti Kuusisto, and Carsten Lutz. Expressive power of graph transformers via logic. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 19569–19579, 2026.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36:57125–57211, 2023.

Rajendra Bhatia. *Matrix analysis*. Springer Science & Business Media, 2013.

Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In *International conference on machine learning*, pp. 864–873. PMLR, 2020.

Chaoqi Chen, Yushuang Wu, Qiyuan Dai, Hong-Yu Zhou, Mutian Xu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. A survey on graph neural networks and graph transformers in computer vision: A task-oriented perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10297–10318, 2024.

Jean Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *8th International Conference on Learning Representations, ICLR 2020*, 2020.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International conference on machine learning*, pp. 2793–2803. PMLR, 2021.

Iryna Gurevych, Michael Kohler, and Gözde Gül Şahin. On the rate of convergence of a classifier based on a transformer encoder. *IEEE Transactions on Information Theory*, 68(12):8139–8155, 2022.

Tiantian He, Yang Liu, Yew-Soon Ong, Xiaohu Wu, and Xin Luo. Polarized message-passing in graph neural networks. *Artificial Intelligence*, 331:104129, 2024.

Yinan Huang, William Lu, Joshua Robinson, Yu Yang, Muhan Zhang, Stefanie Jegelka, and Pan Li. On the stability of expressive positional encodings for graphs. In *International Conference on Learning Representations*, volume 2024, pp. 39745–39774, 2024.

Md Shamim Hussain, Mohammed J Zaki, and Dharmashankar Subramanian. Global self-attention as a replacement for graph convolution. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 655–665, 2022.

Haotian Jiang and Qianxiao Li. Approximation rate of the transformer architecture for sequence modeling. *Advances in Neural Information Processing Systems*, 37:68926–68955, 2024.

Tokio Kajitsuka and Issei Sato. Are transformers with one layer self-attention using low-rank weight matrices universal approximators? In *International Conference on Learning Representations*, volume 2024, pp. 35177–35205, 2024.

Jinwoo Kim, Dat Nguyen, Seonwoo Min, Sungjun Cho, Moontae Lee, Honglak Lee, and Seunghoon Hong. Pure transformers are powerful graph learners. *Advances in Neural Information Processing Systems*, 35:14582–14595, 2022.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Anastasis Kratsios, Behnoosh Zamanlooy, Tianlin Liu, and Ivan Dokmanić. Universal approximation under constraints is possible with transformers. In *International Conference on Learning Representations*.

Devin Kreuzer, Dominique Beaini, Will Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking graph transformers with spectral attention. *Advances in neural information processing systems*, 34:21618–21629, 2021.

Hongkang Li, Meng Wang, Tengfei Ma, Sijia Liu, Zaixi Zhang, and Pin-Yu Chen. What improves the generalization of graph transformers? A theoretical dive into the self-attention and positional encoding. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28784–28829. PMLR, 21–27 Jul 2024a.

Jianfei Li, Han Feng, and Ding-Xuan Zhou. Approximation analysis of cnns from a feature extraction view. *Analysis and Applications*, 22(03):635–654, 2024b.

Jianfei Li, Ruigang Zheng, Han Feng, Ming Li, and Xiaosheng Zhuang. Permutation equivariant graph framelets for heterophilous graph learning. *IEEE Transactions on neural networks and learning systems*, 35(9):11634–11648, 2024c.

Jianfei Li, Han Feng, and Ding-Xuan Zhou. Convergence analysis for deep sparse coding via convolutional neural networks. *Analysis and Applications*, pp. 1–41, 2026.

Shanda Li, Xiangning Chen, Di He, and Cho-Jui Hsieh. Can vision transformers perform convolution? *arXiv preprint arXiv:2111.01353*, 2021.

Jianqing Liang, Min Chen, Xinkai Wei, and Jiye Liang. SCGT: Towards scalable and comprehensive graph transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026, DOI: 10.1109/TPAMI.2026.3682858.

Grégoire Mialon, Dexiong Chen, Margot Selosse, and Julien Mairal. Graphit: Encoding graph structure in transformers. *arXiv preprint arXiv:2106.05667*, 2021.

Luis Müller and Christopher Morris. Aligning transformers with weisfeiler-leman. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36654–36704. PMLR, 21–27 Jul 2024.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *Neural Information Processing Systems (NeurIPS)*, pp. 21263–21274, 2020.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3): 127–239, 2014.

Jinyoung Park, Seongjun Yun, Hyeonjin Park, Jaewoo Kang, Jisu Jeong, Kyung-Min Kim, Jung-Woo Ha, and Hyunwoo J Kim. Deformable graph transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5385–5396, 2025.

Ahsan Shehzad, Feng Xia, Shagufta Abid, Ciyuan Peng, Shuo Yu, Dongyu Zhang, and Karin Verspoor. Graph transformers: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2026, DOI: 10.1109/TNNLS.2025.3646122.

Shokichi Takakura and Taiji Suzuki. Approximation and estimation ability of transformers for sequence-to-sequence functions with infinite dimensional input. In *International Conference on Machine Learning*, pp. 33416–33447. PMLR, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Mingze Wang and Weinan E. Understanding the expressive power and mechanisms of transformer for sequence modeling. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 25781–25856. Curran Associates, Inc., 2024. doi: 10.52202/079017-0811.

Hyowon Wi, Jeongwhan Choi, and Noseong Park. Learning advanced self-attention for linear transformers in the singular value domain. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25, 2025. ISBN 978-1-956792-06-5. doi: 10.24963/ijcai.2025/730. URL https://doi.org/10.24963/ijcai.2025/730.

Qitian Wu, Wentao Zhao, Chenxiao Yang, Hengrui Zhang, Fan Nie, Haitian Jiang, Yatao Bian, and Junchi Yan. Sgformer: Simplifying and empowering transformers for large-graph representations. *Advances in neural information processing systems*, 36:64753–64773, 2023.

Guangrui Yang, Jianfei Li, Ming Li, Han Feng, and Ding-Xuan Zhou. Bridging smoothness and approximation: Theoretical insights into over-smoothing in graph neural networks. *arXiv preprint arXiv:2407.01281*, 2024.

Kaixuan Yao, Jiye Liang, Jianqing Liang, Ming Li, and Feilong Cao. Multi-view graph convolutional networks with attention mechanism. *Artificial Intelligence*, 307:103708, 2022.

Gilad Yehudai, Clayton Sanford, Maya Bechler-Speicher, Orr Fischer, Ran Gilad-Bachrach, and Amir Globerson. Depth-width tradeoffs for transformers on graph tasks. *Advances in Neural Information Processing Systems*, 38:22949–22981, 2026.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34:28877–28888, 2021.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? *ArXiv*, abs/1912.10077, 2019.

Chulhee Yun, Yin-Wen Chang, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. O(n) connections are expressive enough: Universal approximability of sparse transformers. *Advances in Neural Information Processing Systems*, 33:13783–13794, 2020.

Haiteng Zhao, Shuming Ma, Dongdong Zhang, Zhi-Hong Deng, and Furu Wei. Are more layers beneficial to graph transformers? *arXiv preprint arXiv:2303.00579*, 2023.

Cai Zhou, Rose Yu, and Yusu Wang. On the theoretical expressive power and the design space of higher-order graph transformers. In *International Conference on Artificial Intelligence and Statistics*, pp. 2179–2187. PMLR, 2024.

Jun Zhou, Guojie Li, Chang Xu, and Jianxin Liu. Dirichlet energy of graph neural networks: A unified view of expressiveness. In *International Conference on Machine Learning (ICML)*, pp. 12677–12687, 2021.

Wenhao Zhu, Tianyu Wen, Guojie Song, Liang Wang, and Bo Zheng. On structural expressive power of graph transformers. In *Proceedings of the 29th acm sigkdd conference on knowledge discovery and data mining*, pp. 3628–3637, 2023.

A Deferred Proofs

A.1 Proof of Theorem 1

Proof of Theorem 1. From equation 1,

$$V^\top \bar{x}_i' = V^\top \sum_{j=1}^{N} \langle Qx_i, Kx_j \rangle V x_j$$
$$= \sum_{j=1}^{N} \sum_{s=1}^{\ell} h_{q,s}(x_i) h_{k,s}(x_j) V^\top V x_j.$$

Using the identity $V^\top V x_j = \sum_{t=1}^{d'} \langle x_j, v_t \rangle v_t$, we obtain

$$V^\top \bar{x}_i' = \sum_{j=1}^{N} \sum_{s=1}^{\ell} h_{q,s}(x_i) h_{k,s}(x_j) \sum_{t=1}^{d'} \langle x_j, v_t \rangle v_t$$
$$= \sum_{s=1}^{\ell} \sum_{t=1}^{d'} \left(\sum_{j=1}^{N} h_{k,s}(x_j) \langle x_j, v_t \rangle \right) \big(h_{q,s}(x_i) v_t \big). \tag{7}$$

By the definition of $\phi_{s,t}(\widetilde{X})$,

$$\big\langle \widetilde{X}, \phi_{s,t}(\widetilde{X}) \big\rangle = \sum_{j=1}^{N} \langle x_j, h_{k,s}(x_j) v_t \rangle = \sum_{j=1}^{N} h_{k,s}(x_j) \langle x_j, v_t \rangle. \tag{8}$$

Moreover, the i-th block of $\psi_{s,t}(\widetilde{X})$ is exactly $h_{q,s}(x_i) v_t$. Substituting equation 8 into equation 7 yields, for every i,

$$V^\top \bar{x}_i' = \sum_{s=1}^{\ell} \sum_{t=1}^{d'} \big\langle \widetilde{X}, \phi_{s,t}(\widetilde{X}) \big\rangle \, \psi_{s,t}(\widetilde{X})_i, \tag{9}$$

where $\psi_{s,t}(\widetilde{X})_i$ denotes the i-th block of $\psi_{s,t}(\widetilde{X})$. Concatenating these identities over $i = 1, \ldots, N$ proves equation 2. □

A.2 Proof of Theorem 2

Proof of Theorem 2. For $j \in \mathcal{N}_i$, let

$$s_{ij} = \langle Qx_i, Kx_j \rangle = \sum_{s=1}^{\ell} h_{q,s}(x_i) h_{k,s}(x_j).$$

Using the exponential-kernel expansion,

$$\exp(s_{ij}) = \sum_{\nu \in \mathbb{N}_0^\ell} \frac{h_q^\nu(x_i) h_k^\nu(x_j)}{\nu!}.$$

Therefore, using $V^\top V x_j = \sum_{t=1}^{d'} \langle x_j, v_t \rangle v_t$,, we obtain

$$V^\top \bar{x}_i^{\mathrm{sm}} = \sum_{j \in \mathcal{N}_i} \frac{\exp(s_{ij})}{\mathcal{Z}_i} V^\top V x_j$$
$$= \sum_{j \in \mathcal{N}_i} \frac{1}{\mathcal{Z}_i} \left(\sum_{\nu \in \mathbb{N}_0^\ell} \frac{h_q^\nu(x_i) h_k^\nu(x_j)}{\nu!} \right) \left(\sum_{t=1}^{d'} \langle x_j, v_t \rangle v_t \right).$$

Rearranging the absolutely convergent sums gives

$$V^\top \bar{x}_i^{\mathrm{sm}} = \sum_{\nu \in \mathbb{N}_0^\ell} \sum_{t=1}^{d'} \frac{h_q^\nu(x_i)}{\nu! \mathcal{Z}_i} \left(\sum_{j \in \mathcal{N}_i} h_k^\nu(x_j) \langle x_j, v_t \rangle \right) v_t.$$

By the definition of $\phi^{\mathrm{sm}}_{i,\nu,t}(\tilde{X})$,

$$\left\langle \tilde{X}, \phi^{\mathrm{sm}}_{i,\nu,t}(\tilde{X}) \right\rangle = \sum_{j \in \mathcal{N}_i} h^{\nu}_k(x_j) \langle x_j, v_t \rangle.$$

Moreover, the i-th block of $\psi^{\mathrm{sm}}_{i,\nu,t}(\tilde{X})$ is exactly $\frac{h^{\nu}_q(x_i)}{\nu! \mathcal{Z}_i} v_t$, and all other blocks are zero. Hence, after concatenating the identities over $i = 1, \ldots, N$, we obtain

$$\tilde{Y}^{\mathrm{sm}} = \sum_{i=1}^{N} \sum_{\nu \in \mathbb{N}^{\ell}_0} \sum_{t=1}^{d'} \left\langle \tilde{X}, \phi^{\mathrm{sm}}_{i,\nu,t}(\tilde{X}) \right\rangle \psi^{\mathrm{sm}}_{i,\nu,t}(\tilde{X}),$$

as claimed. □

A.3 Proof of Corollary 1

Proof of Corollary 1. We explicitly construct the required matrices Q, K, V. Since $z_i = [x_i^\top, e_i^\top]^\top$, we define Q and K to act only on the positional part, ignoring the original features. This is achieved by setting the first d columns of Q and K to zero:

$$Q z_i = \widetilde{Q}\, e_i, \qquad K z_j = \widetilde{K}\, e_j,$$

with $\widetilde{Q}, \widetilde{K} \in \mathbb{R}^{\ell \times N}$.

Let $\ell = \mathrm{rank}(A)$. By the singular value decomposition,

$$A = \sum_{s=1}^{\ell} \sigma_s\, u_s v_s^\top,$$

where $\sigma_s > 0$ and $u_s, v_s \in \mathbb{R}^N$ are the left and right singular vectors. Define the matrices

$$\widetilde{Q}_{s,i} := \sqrt{\sigma_s}\, u_{s,i}, \qquad \widetilde{K}_{s,j} := \sqrt{\sigma_s}\, v_{s,j},$$

for $s = 1, \ldots, \ell$ and $i, j = 1, \ldots, N$. Then the inner product of the query and key vectors is

$$\langle Q z_i, K z_j \rangle = \sum_{s=1}^{\ell} \widetilde{Q}_{s,i} \widetilde{K}_{s,j} = \sum_{s=1}^{\ell} \sigma_s\, u_{s,i} v_{s,j} = A_{ij}.$$

Now define the value matrix V to act only on the original feature part x_j, discarding the positional encoding, and applying the desired projection P:

$$V z_j = W x_j.$$

This is achieved by setting the first d columns of V to P and the remaining N columns to zero.

Substituting these constructions into the linearized aggregation rule gives

$$\bar{x}'_i = \sum_{j=1}^{N} \langle Q z_i, K z_j \rangle\, (V z_j) = \sum_{j=1}^{N} A_{ij}\, W x_j.$$

Stacking the outputs for all nodes $i = 1, \ldots, N$ yields

$$\bar{X}' = \begin{bmatrix} \bar{x}'_1 \\ \vdots \\ \bar{x}'_N \end{bmatrix} = \begin{bmatrix} \sum_j A_{1j} W x_j \\ \vdots \\ \sum_j A_{Nj} W x_j \end{bmatrix} = A X W^\top.$$

This is precisely the desired graph convolution. Applying a pointwise nonlinearity σ afterwards recovers the standard layer $\sigma(AXW^\top)$. □

A.4 Proof of Theorem 3

Proof of Theorem 3. The smooth part of $\mathcal{J}$ is

$$f(Z) = \frac{1}{2}\|XV^\top - D_\theta Z\|_F^2,$$

and

$$\nabla f(Z) = D_\theta^\top (D_\theta Z - XV^\top).$$

Therefore a proximal-gradient step Parikh & Boyd (2014) is

$$Z^{k+1} = \mathrm{prox}_{\eta_k \alpha \|\cdot\|_{2,1}} \left(Z^k - \eta_k D_\theta^\top (D_\theta Z^k - XV^\top) \right).$$

Equivalently,

$$\widetilde{Z}^{k+1} = Z^k + \eta_k D_\theta^\top (XV^\top - D_\theta Z^k).$$

The proximal map of the row-group norm is the group soft-thresholding operator Parikh & Boyd (2014),

$$\mathrm{prox}_{\eta_k \lambda \|\cdot\|_{2,1}}(U)_{:,j} = \left(1 - \frac{\eta_k \alpha}{\|U_{:,j}\|_2}\right)_+ U_{:,j}.$$

This gives the claimed recurrence. The GT and GCN interpretations follow by choosing D_θ^k from the adaptive attention family or from the fixed topology-driven family, respectively.

□

A.5 Proof of Theorem 4

Proof of Theorem 4. To establish the lower bound on the reconstruction gap, it suffices to derive a lower bound for $\tilde{\Delta} := \mathbb{E}[\ell(Z_A)] - \mathbb{E}[\ell(Z_{\tilde{A}})]$, where $\tilde{A} = P\mathrm{Diag}(\theta_1, \ldots, \theta_N)P^\top$ and $P, \theta_i, i \in [N]$, are trainable parameters. In addition we assume that P is orthonormal. This reduction is validated by Corollary 1, since GT can be reduced to any graph convolution filter $\tilde{A}$, which ensures that $\Delta \geq \tilde{\Delta}$. We can also write the notation more compactly as $D = U\mathrm{Diag}(d_1, \ldots, d_N)U^\top$, where $d_i = \lambda_i$ and $U = H$ when $D = A$, and $d_i = \theta_i$ and $U = P$ when $D = \tilde{A}$.

Notice that the minimizer of equation 3 w.r.t $D_\theta := D$ is $Z_D := Z_\theta = (D^\top D + \alpha I)^{-1} D^\top X$. Using $X = X^* + \eta$ and $\mathbb{E}[\eta] = \mathbf{0}$:

$$\begin{aligned} \mathbb{E}[\ell(Z_D)] &= \mathbb{E}[\|X^* - DZ_D\|_F^2] \\ &= \|X^* - D(D^\top D + \alpha I)^{-1} D^\top X^*\|_F^2 \\ &\quad + \mathbb{E}\left[\|D(D^\top D + \alpha I)^{-1} D^\top \eta\|_F^2\right]. \end{aligned} \tag{10}$$

Since $\mathbb{E}[\eta\eta^\top] = \sigma^2 d I_N$, we have:

$$\begin{aligned} &\mathbb{E}\left[\|D(D^\top D + \alpha I)^{-1} D^\top \eta\|_F^2\right] \\ &= \mathbb{E}\,\mathrm{tr}\left(D(D^\top D + \alpha I)^{-1} D^\top \eta\eta^\top D(D^\top D + \alpha I)^{-1} D^\top\right) \\ &= \mathbb{E}\,\mathrm{tr}\left(\left(D(D^\top D + \alpha I)^{-1} D^\top\right)^2 \eta\eta^\top\right) \\ &= \sigma^2 d \cdot \mathrm{tr}\left(\left(D(D^\top D + \alpha I)^{-1} D^\top\right)^2\right), \end{aligned} \tag{11}$$

Let $G := D(D^\top D + \alpha I)^{-1} D^\top = U\mathrm{Diag}(b_1, \ldots, b_n)U^\top$ where $b_i = \frac{d_i^2}{d_i^2 + \alpha}$. The bias term of equation 10 is given by

$$\begin{aligned} \|X^* - GX^*\|_F^2 &= \sum_{i=1}^N (1 - b_i)^2 \|U_{:,i}^\top X^*\|_2^2 \\ &= \sum_{i=1}^N \left(\frac{\alpha}{d_i^2 + \alpha}\right)^2 \|U_{:,i}^\top X^*\|_2^2. \end{aligned}$$

while the variance term of equation 10, which is simplified to equation 11, is given by

$$\sigma^2 d \cdot \mathrm{tr}(G^2) = \sigma^2 d \sum_{i=1}^{N} b_i^2 = \sigma^2 d \sum_{i=1}^{N} \left(\frac{d_i^2}{d_i^2 + \alpha} \right)^2 .$$

Combining both terms into equation 10 gives the expression of $\mathbb{E}[\ell(Z_D)]$

$$\mathbb{E}[\ell(Z_D)] = \sum_{i=1}^{N} \left(\frac{\alpha}{d_i^2 + \alpha} \right)^2 \|U_{:,i}^\top X^*\|_2^2 + \sigma^2 d \sum_{i=1}^{N} \left(\frac{d_i^2}{d_i^2 + \alpha} \right)^2 . \tag{12}$$

In the case where $D = \tilde{A}$, we can additionally minimize θ_i and thereby simplify the expression for $\mathbb{E}[\ell(Z_{\tilde{A}})]$. Define $a_i := \|P_{:,i}^\top X^*\|_2^2$ and $x := \theta_i^2 \geq 0$. Notice that $\mathbb{E}[\ell(Z_D)]$ can be written as a sum over i, i.e., it is separable with respect to i. In the adaptive filter setting, the optimal value of θ_i is obtained by minimizing $f_i(x) = \left(\frac{\alpha}{x+\alpha}\right)^2 a_i + \sigma^2 d \left(\frac{x}{x+\alpha}\right)^2$. Taking derivative, we obtain

$$f_i'(x) = \frac{-2\alpha^2 a_i + 2\sigma^2 d\alpha x}{(x+\alpha)^3}.$$

Setting $f_i'(x) = 0$ gives $x_i^* = \frac{\alpha a_i}{d\sigma^2}$. It is easy to verify that x_i^* is the global minimum. Substituting $x_i^* = \frac{\alpha a_i}{d\sigma^2}$ into f_i, the optimal risk for each i is

$$\begin{aligned} f_i(x_i^*) &= \frac{d^2\sigma^4 a_i}{(\alpha d\sigma^2 + \alpha a_i)^2} \cdot \alpha^2 + d\sigma^2 \cdot \frac{\alpha^2 a_i^2}{(\alpha d\sigma^2 + \alpha a_i)^2} \\ &= \frac{\sigma^2 d\alpha a_i}{\alpha\sigma^2 d + \alpha a_i} = \frac{\sigma^2 d a_i}{\sigma^2 d + a_i}. \end{aligned} \tag{13}$$

Therefore, using equation 12 for $D = A$ and equation 13 for $D = \tilde{A}$, we derive

$$\begin{aligned} \tilde{\Delta} &= \sum_{i=1}^{N} \left[\left(\frac{\alpha}{\lambda_i^2 + \alpha} \right)^2 \|H_{:,i}^\top X^*\|_2^2 + \sigma^2 d \left(\frac{\lambda_i^2}{\lambda_i^2 + \alpha} \right)^2 \right] \\ &\quad - \sum_{i=1}^{N} \frac{\sigma^2 d a_i}{\sigma^2 d + a_i} \\ &= \sum_{i=1}^{N} \left[\frac{\alpha^2 \|H_{:,i}^\top X^*\|_2^2 + \sigma^2 d \lambda_i^4}{(\lambda_i^2 + \alpha)^2} - \frac{\sigma^2 d a_i}{\sigma^2 d + a_i} \right]. \end{aligned}$$

Recall that $\Delta \geq \tilde{\Delta}$ and P can be chosen arbitrarily. Together with the above estimate of $\tilde{\Delta}$, we complete the proof. □

A.6 Proof of Corollary 3

Proof of Corollary 3. Let us consider the optimization problem

$$\min_{\{P_{:,i}\}_{i=1}^N} \sum_{i=1}^{N} \frac{\sigma^2 d \, \|P_{:,i}^\top X^*\|_2^2}{\sigma^2 d + \|P_{:,i}^\top X^*\|_2^2} \qquad \text{s.t.} \quad P^\top P = I_N. \tag{14}$$

Define $B = X^*(X^*)^\top \in \mathbb{R}^{N\times N}$. Since B is symmetric and positive semidefinite, it admits an eigendecomposition $B = P_B \cdot \mathrm{Diag}(\mu_1, \ldots, \mu_N) \cdot P_B^\top$ with $\mu_1 \geq \cdots \geq \mu_N \geq 0$. For any $P_{:,i}$, we have

$$\|P_{:,i}^\top X^*\|_2^2 = P_{:,i}^\top X^* (X^*)^\top P_{:,i} = P_{:,i}^\top B P_{:,i}.$$

The objective function can therefore be rewritten as

$$F(P) = \sum_{i=1}^{N} f\big(P_{:,i}^\top B P_{:,i}\big), \quad \text{where} \quad f(t) = \frac{\sigma^2 d\, t}{\sigma^2 d + t}, \; t \geq 0.$$

The function f is smooth on $[0, \infty)$ and strictly concave because

$$f''(t) = -\frac{2\sigma^4 d^2}{(\sigma^2 d + t)^3} < 0 \quad \text{for all } t \geq 0.$$

The feasible set $\{P \in \mathbb{R}^{N\times N} : P^\top P = I_N\}$ is compact and F is continuous, so a global minimum exists by the extreme value theorem. To characterize the minimum we use majorisation.

For any orthogonal matrix P, we collect the diagonal entries of $P^\top BP$ into the vector

$$d(P) = \big(P_{:,1}^\top BP_{:,1},\ \ldots,\ P_{:,N}^\top BP_{:,N}\big)^\top.$$

Let $d^\downarrow(P)$ be its decreasing rearrangement. The Schur's Theorem asserts that $d^\downarrow(P)$ is majorised by the eigenvalue vector $\mu = (\mu_1, \ldots, \mu_N)^\top$:

$$d^\downarrow(P) \ \prec\ \mu.$$

More explicitly,

$$\sum_{i=1}^{k} d_i^\downarrow(P) \leq \sum_{i=1}^{k} \mu_i \quad k \in [N],$$
$$\sum_{i=1}^{N} d_i(P) = \sum_{i=1}^{N} \mu_i.$$

Because f is concave, (Bhatia, 2013, Theorem II.3.1) implies

$$\sum_{i=1}^{N} f(\mu_i) \ \leq\ \sum_{i=1}^{N} f\big(d_i^\downarrow(P)\big) \ =\ \sum_{i=1}^{N} f\big(P_{:,i}^\top AP_{:,i}\big) = F(P).$$

Hence $F(P) \geq \sum_{i=1}^{N} f(\mu_i)$ for every feasible P.

This lower bound is attainable. Choose $P = P_B = (p_1^B, \ldots, p_N^B)$, where the eigenvector p_i^B corresponds to μ_i. Then $(p_i^B)^\top Bp_i^B = \mu_i$ and

$$F(P_B) = \sum_{i=1}^{N} f(\mu_i). \tag{15}$$

Therefore, the minimum value is $\sum_{i=1}^{N} f(\mu_i)$, achieved by any orthogonal matrix whose columns are eigenvectors of B.

Next, we derive a lower bound for the special case $P = H$. Since $H \in \{P : P^\top P = I_N\}$, this bound will be weaker than equation 15, where we optimized over P. Let $a_i := \|H_{:,i}^\top Y\|_2^2$. Then

$$\Delta \geq \sum_{i=1}^{N} \left[\frac{\alpha^2 a_i + \sigma^2 d\lambda_i^4}{(\lambda_i^2 + \alpha)^2} - \frac{\sigma^2 d a_i}{\sigma^2 d + a_i}\right]$$
$$= \sum_{i=1}^{N} \frac{(\alpha a_i - \sigma^2 d\lambda_i^2)^2}{(\lambda_i^2 + \alpha)^2 (a_i + \sigma^2 d)}.$$

Using $(\lambda_i^2 + \alpha)^2 \leq (\lambda_1^2 + \alpha)^2$ and $a_i + \sigma^2 d \leq \|X^*\|_F^2 + \sigma^2 d$, we can further simplify the lower bound as

$$\Delta \geq \frac{1}{(\lambda_1^2 + \alpha)^2 (\|X^*\|_F^2 + \sigma^2 d)} \sum_{i=1}^{N} (\alpha a_i - \sigma^2 d\lambda_i^2)^2. \tag{16}$$

By applying the Cauchy–Schwarz inequality to the vectors $(\lambda_i^2)_{i=1}^N$ and $\left(\alpha a_i - \sigma^2 d\lambda_i^2\right)_{i=1}^N$, we can rewrite and simplify the summation term as follows:

$$\begin{aligned}\sum_{i=1}^{N}(\alpha a_i - \sigma^2 d\lambda_i^2)^2 &\geq \frac{1}{\|A^2\|_F^2}\left(\sum_{i=1}^{N}\lambda_i^2(\alpha a_i - \sigma^2 d\lambda_i^2)\right)^2 \\ &= \frac{1}{\|A^2\|_F^2}\left(\alpha\sum_{i=1}^{N}\lambda_i^2 a_i - \sigma^2 d\sum_{i=1}^{N}\lambda_i^4\right)^2 \\ &\geq \frac{1}{\|A^2\|_F^2}\left(\alpha\|AX^*\|_F^2 - \sigma^2 d\|A^2\|_F^2\right)^2.\end{aligned} \tag{17}$$

By inserting equation 17 into equation 16, we obtain the claimed lower bound

$$\Delta \geq \frac{(\alpha\|AX^*\|_F^2 - \sigma^2 d\|A^2\|_F^2)^2}{(\lambda_1^2+\alpha)^2(\|X^*\|_F^2 + \sigma^2 d)\|A^2\|_F^2}$$

This concludes the proof. □

B Additional Interpretation of the Reconstruction Bound

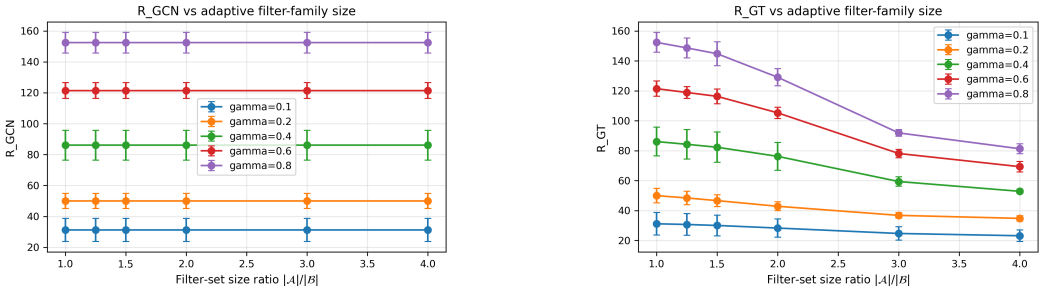
The two sums in Corollary 3 are indexed independently: the first uses the eigenvalues of the fixed filter and the projections of X^* onto its eigenvectors, whereas the second uses the eigenvalues of $X^*(X^*)^\top$. They therefore compare processing in the fixed-filter basis with processing in the signal's principal basis.

When $\sigma = 0$, the simplified lower bound becomes

$$\alpha\|AX^*\|_F^4(\lambda_1^2+\alpha)^{-2}\|X^*\|_F^{-2}\|A^2\|_F^{-2}.$$

Because $\|AX^*\|_F^2 = \sum_i \lambda_i^2\|H_{:,i}^\top X^*\|_2^2$, directions with both large filter magnitude and large signal projection dominate. The bound is one-sided, depends on the unknown clean signal, and may loosen at each relaxation step. Consequently, a zero lower bound does not imply a zero reconstruction gap, and reconstruction risk need not predict downstream-task performance.

C Additional Experimental Details



(a) Best fixed-family risk.

(b) Best adaptive-family risk.

Figure 5: Experiment 3: decomposition of the risk-gap curve. The fixed-family risk remains unchanged as $|\mathcal{A}|/|\mathcal{B}|$ increases, while the adaptive-family risk decreases when additional spectral filters are introduced. This confirms that the gap in Fig. 4 is caused by the enlarged adaptive filter family rather than by changes to the data or evaluation protocol.