DUAL-FREQUENCY DECOUPLING FOR EXPLORING INHERENT STRUCTURAL AND TOPOLOGICAL FEATURES IN HOMOGENEOUS GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

While graph transformer methods effectively enlarge the receptive field, the complexity of homogeneous graph data compromises global attention's ability to capture local structural features. This limitation introduces task-irrelevant long-range topological information, resulting in task representation inaccuracies. To address this challenge, we propose a high- and low-frequency decoupled adaptive constraint method to explicitly capture structural and topological features. First, a dual-frequency decoupling mechanism is designed in the feature capture phase to refine long-range topological information and better distinguish features. Subsequently, residual attention gating augmented by long-range prior knowledge is introduced in the feature integration phase to ensure coherent integration. Finally, multi-scale consistency constraints are proposed in the representation phase to align the integrated representation with the mapping of long-range topological information, ensuring robust training. Our method was benchmarked on six homogeneous datasets, outperforming state-of-the-art methods by 2.15% and 2% on the Cora and Citeseer datasets, respectively.

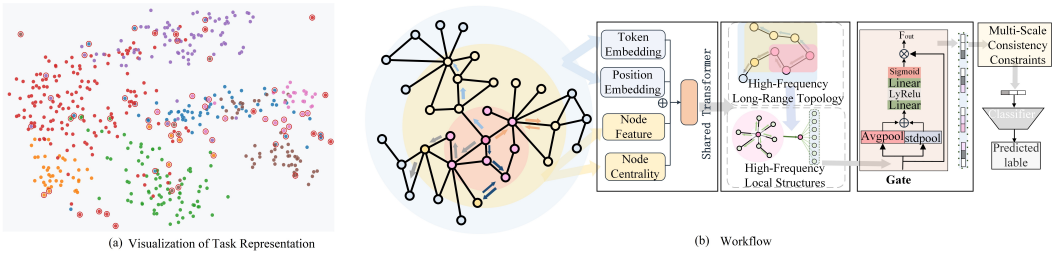
1 INTRODUCTION

Graphs provide a natural representation for relational systems, including citation networks, recommendation systems, transportation networks, and molecular structures (Feng et al., 2022; Choi et al., 2026). While, representation learning faces significant challenges in capturing the inherent complexity and intricate features of various graph structures. The exploration is conducted from the following three aspects. Graph neural network (GNN) approaches emerged as a dominant paradigm for solving graph-based representation tasks (Xu & Ding, 2026). Traditional message-passing GNNs aggregate neighboring representations, iteratively updating the model. However, these approaches retain issues, including over-smoothing and over-compression. Graph transformers (Zhang et al., 2026) have demonstrated outstanding performance leveraging their powerful modeling capabilities, making these architectures effective for processing graph structures.

Graph transformer approaches treat the entire input graph as a fully connected entity, wherein individual nodes are interpreted as independent tokens, thereby constructing a unified sequence encompassing all nodes. This alleviates limitations inherent in message-passing neural networks (MPNNs) (Gilmer et al., 2017). While this approach extends global perceptual insight, promoting node representation learning, it also introduces irrelevant long-range topological noise and causes structural information loss. These limitations potentially hinder a comprehensive understanding of whole-graph structures. In recent years, researchers have developed token-based methods (Zhou et al., 2025; Chen et al., 2023; Xing et al., 2025)(Kreuzer et al. 2021; Zhao et al. 2021; Zhang et al. 2022) that generate a sequence for each node, outperforming traditional graph transformers. However, these approaches encounter challenges when dealing with varied graph structures: excessively relying on multiscale token modeling while lacking exploration of long-range topological capabilities.

Recent graph transformer methods (Xing et al., 2025; Zhou et al., 2025) have achieved competitive results on homogeneous graphs. However, their overreliance on multiscale token structural modeling causes limitations in task representation. Specifically, they neglect both the structural relationships

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.



(a) Visualization of Task Representation

(b) Workflow

Figure 1: Task Representation Results and Workflow. (a) The node classification task representation results. (b) The workflow of the method proposed in this paper.

within node neighborhoods and the long-range topological information between neighborhoods, limiting their performance on diverse graph types like heterogeneous graphs. In Figure 1, the task representation capability of graph transformer models declines after ablating the long-range represented expression. To overcome these limitations, the high- and low-frequency decoupling strategy must be optimized in two key aspects: 1) **Fine-grained local structure:** Accounting for the inherent diversity and complexities of graph structures, token representations should focus more intently, effectively capturing intricacies within neighborhoods and topological information. 2) **Long-range topological integrity:** To obtain a more comprehensive learning of graph structures, long-range topology should provide approximate full coverage, capturing both global and local information to improve task representation capabilities.

Thus, we propose a novel high- and low-frequency decoupling method, its core achieving multi-scale token decomposition into dual-frequency signals and training predictors identifying fine-grained features. Figure X shows the dual-frequency decoupling mechanism decomposing the token signal encoded from a shared transformer, separating it into local structural attribute features and long-range topological dependency information. Second, residual attention gating allows local queries to retrieve relevant long-range topological information, enabling consistent feature representation via long-range prior knowledge. Finally, the multiscale consistency constraint aligns with bounded coherence perception to produce the final representation. The main contributions of this work are as follows:

- We propose a high- and low-frequency decoupling mechanism to capture finer-grained node features via dual-frequency encoding, mitigating overreliance on token representations and improving representation accuracies.
- We introduce residual attention gating in conjunction with long-range prior knowledge, enabling global and local adaptation fusion while preventing confusion between dual-frequency representations, thereby enhancing task representation capabilities.
- We design multi-scale consistency constraints to enable the model's bounded, consistent representation of graph structural complexity and long-range topology, improving model stability.

2 Related Work

2.1 Message-Passing GNNs and Graph Transformers

Graph neural networks (GNNs) established message passing as the dominant way to encode relational structure. A message-passing layer aggregates features from a node's neighbors and applies a node-wise update, providing a strong locality inductive bias and making models such as GCN, GraphSAGE, and GAT effective on citation and social graphs (Kipf & Welling, 2017; Hamilton et al., 2017; Veličković et al., 2018). However, repeated aggregation can drive nearby node states toward an indistinguishable subspace (over-smoothing) (Chen et al., 2020), while a fixed-width message must compress an exponentially growing number of paths (oversquashing) (Alon & Yahav, 2021). Studies such as FAGCN and GPR-GNN (Bo et al., 2021; Chien et al., 2021) point out that low-pass propagation is not suitable for all graph data, and that incorporating high-frequency components along with dataset-adaptive modules can improve model performance. These approaches

rely on graph filters and propagation coefficients to achieve multi-scale structural decomposition, enabling the simultaneous modeling of graph signals at different frequencies. However, they remain confined within the message-passing framework and fail to construct explicit node-centered token sequences for flexibly modeling long-range interactions between nodes.

Graph Transformers (GTs) address the receptive-field bottleneck by replacing local aggregation with multi-head attention over a broader set of nodes. Early GTs introduced graph-specific positional encodings (Dwivedi & Bresson, 2021; Deng et al., 2024); Graphormer enriched attention with centrality, spatial, and edge encodings (Ying et al., 2021); and sparse variants such as Exphormer and VCR-Graphormer t (Shirzad et al., 2023; Fu et al., 2024) extend the contextual scope while reducing the computational overhead of fully connected attention, shifting the research focus from identifying reachable neighbors to screening for effective node interactions. However, global attention introduces a new issue of "over-globalization": irrelevant distant nodes can interfere with meaningful local information, while attention and positional encodings tend to conflate local topology with long-range context. The work of Xing et al. (Xing et al., 2024) makes this issue explicit: global access is not equivalent to useful access, and a single interaction pattern is unlikely to be optimal across graphs or nodes. Unified-mask work reaches a related conclusion by interpreting local, cluster-level, and global GTs as complementary attention masks whose receptive-field size and label consistency jointly affect classification (Xing et al., 2025).

In summary, while GTs alleviate the locality constraints of GNNs, they still struggle to balance fine-grained local neighborhood information with task-relevant long-range dependencies. This inherent trade-off serves as the starting point for this research.

2.2 Tokenized and Hierarchical Graph Transformers

Token-based GTs replace unrestricted graph-wide attention with a node-centered sequence. They construct structural tokens from propagation states, sampled paths, virtual connections, or graph partitions and then apply a Transformer to those tokens. NAGphormer uses Hop2Token to turn successive neighborhood-aggregation states into a sequence (Chen et al., 2023). Exphormer uses sparse expander-style and virtual-global connections (Shirzad et al., 2023), VCR-Graphormer introduces virtual connections for mini-batch training (Fu et al., 2024), Polynormer uses polynomial-expressive attention in linear time (Deng et al., 2024), and cluster-wise GTs add an intermediate interaction scale (Huang et al., 2024). These methods improve scalability and make the receptive field more controllable, but each tokenization or mask remains a design choice that privileges particular interactions.

The limitation becomes sharper when token sequences are treated as one undifferentiated evidence pool. Hop tokens summarize local neighborhoods efficiently but may compress relations among nodes at the same hop and omit dependencies beyond the selected radius. Path or virtual-connection tokens enlarge coverage but can introduce redundant or task-irrelevant topology. Hierarchical-mask analyses likewise show that local, cluster, and global interactions are complementary and that naive averaging can be worse than selecting the appropriate interaction level (Xing et al., 2025). Tokenization therefore addresses the cost of global attention but does not automatically solve the representation problem identified in the Introduction: fine-grained local structure and long-range topological integrity must be preserved simultaneously and combined selectively.

Tokenphormer advances this line by generating multiple structure-aware tokens from mixed random walks, hop neighborhoods, and a serialized-graph prior, and by jointly feeding them to a Transformer (Zhou et al., 2025). Walk tokens provide varied receptive fields, hop tokens provide local coverage, and the SGPM token supplies graph-level context. Nevertheless, the token types are still exposed to a common Transformer fusion process; the model does not explicitly enforce a highfrequency/low-frequency division, nor does it constrain an unreliable global prior from overriding node-specific local evidence.

To address the two open questions of how local tokens can acquire long-range information without collapsing their structural roles, and how prior alignment can improve topology while preserving local semantics, a viable method is to treat Hop tokens and Walk tokens as high-frequency views of neighborhood variations and low-frequency views of long-range topology. By employing independent branch encoding, cross-frequency attention, and adaptive hierarchical fusion, this approach

clarifies token roles, routing mechanisms, and metric constraints through multi-scale constraints while protecting self-tokens and aligning priors, thereby closing the representation learning loop.

3 METHOD

We present proposed method as a three-stage workflow. High- and low-frequency decoupling first captures long-range topological information and local structure from shared transformer encoders. Next, residual attention gating ensures consistent representation of structural and topological information, generating the representation space. Finally, multi-scale consistency constraints aligns aligns and optimizes the representation space. For simplicity, we focus on homogeneous graph data node classification with extensions to heterogeneous graph data. Based on this, we propose the following informal assumptions:

Assumption (task-conditional complementarity). For a nonzero-measure subset of nodes, neither frequency view is sufficient by itself. With Y_i denoting the label, we assume

$$I(Y_i; \ell_i \mid \widehat{h}_i) > 0, \qquad I(Y_i; \widehat{h}_i \mid \ell_i) > 0, \tag{1}$$

where $\widehat{h}_i$ and ℓ_i are the summaries constructed below. The assumption only guarantees strict complementarity between high and low frequencies, providing the foundational mathematical conditions for the following assumption.

Assumption (bounded gated prior fusion). All token summaries and intermediate residuals are finite on the representation domain; the residual map and gate networks are locally Lipschitz; and every adaptive coefficient is produced by a sigmoid gate. The Long-range topological prior may contain node-irrelevant components, so it is amplified only through a clipped cosine agreement with the current fused representation. This assumption is a sufficient modeling condition: it does not claim that every prior is useful, but guarantees that an incompatible prior cannot cause unbounded amplification. The two assumptions jointly define the regime analyzed below: complementary views provide potentially useful information, while bounded gates determine how that information is admitted.

3.1 HIERARCHICAL DUAL-FREQUENCY TOKEN ENCODING

This stage proposes a Hierarchical Dual-Frequency Token Encoding (HDTE) mechanism, which converts raw Hop- and Walk-tokens into two contextualized streams and passes both streams, rather than a prematurely averaged vector, to adaptive fusion.

For a graph $G = (V, E)$ and target node v_i, tokenization yields a self-token $x_i \in \mathbb{R}^d$, an Long-range topological prior token $s_i \in \mathbb{R}^d$, Hop-tokens $\mathcal{H}_i \in \mathbb{R}^{K_h \times d}$, and Walk-tokens $\mathcal{W}_i \in \mathbb{R}^{K_w \times d}$. We use "high frequency" operationally for short-range connectivity variation and "low frequency" for slowly varying context aggregated by longer walks; this is not an exact graph-Fourier decomposition.

Parallel frequency-specific encoding. The Walk branch performs self-attention inside the trajectory set, while the Hop branch performs the analogous operation inside each local-radius set. Layer normalization and attentive pooling are standard operations shared by both paths, so we denote their outputs compactly by $(\widetilde{\mathcal{W}}_i, \ell_i^0; \mathcal{H}_i^0, h_i^0) = \mathrm{Encode}_{L/H}(\mathcal{W}_i, \mathcal{H}_i)$. Where ℓ_i^0 summarizes long-range reachability and h_i^0 summarizes local structural variation. Separate parameters preserve branch identity and allow the next step to judge branch reliability independently.

Context-conditioned Walk refinement. The self-token and local seed form a query context $c_i = \tanh(W_c[x_i \| h_i^0] + b_c)$. Each Walk-token receives a score combining learned compatibility and cosine agreement, followed by a differentiable soft summary:

$$e_{ij} = \frac{w_e^\top \tanh(W_w w_{ij} + c_i)}{\tau_s} + \gamma_s \cos(\tanh(W_w w_{ij} + c_i), c_i), \ell_i^{\mathrm{soft}} = \sum_j \mathrm{softmax}_j(e_{ij}) w_{ij}. \tag{2}$$

The workflow retains the top-K_s candidates, applies self-attention to this reduced set, and obtains ℓ_i^{top}. A gate interpolates the soft and selected summaries, preserving gradient flow while concentrating capacity on walks compatible with the local context. The sampled summary is blended with

ℓ_i^0 and retained as a residual correction:

$$\Delta_i = \rho_s \exp\left(\kappa_s \cos(\bar{\ell}_i, \ell_i^0)\right)(\bar{\ell}_i - \ell_i^0), \qquad \bar{\ell}_i = b_i \odot \ell_i^{\mathrm{sam}} + (1 - b_i) \odot \ell_i^0. \tag{3}$$

Given $\rho_s = 0$ removes only this refinement, not the low-frequency encoder.

Cross-frequency conditioning. After the Walk stream has been contextualized, the Hop stream queries the complete Walk-token sequence $\widehat{\mathcal{H}}_i = \mathrm{CrossAttn}_C(\mathcal{H}_i, \widetilde{\mathcal{W}}_i)$, $\widehat{h}_i = \mathrm{Pool}_H(\widehat{\mathcal{H}}_i)$. Long-range context therefore disambiguates local patterns, while the unmodified ℓ_i^0 remains available to fusion. The interface is $(\mathcal{H}_i, \mathcal{W}_i) \to (h_i^0, \ell_i^0) \to \ell_i^{\mathrm{sam}} \to (\widehat{h}_i, \ell_i^0)$, explicitly separating evidence construction from evidence selection. HDTE preserves branch identity, refines noisy long-range candidates before fusion, and returns a cross-conditioned local view together with an independent low-frequency reference.

The construction also subsumes either single branch, which yields the following capacity result; its proof is deferred to Appendix A.2.

Theorem 1 (Complementary capacity)**.** *Let $\mathcal{F}_H$, $\mathcal{F}_L$, and $\mathcal{F}_{HL}$ denote the predictor classes induced by the high-frequency branch, the low-frequency branch, and the joint encoder, respectively. If the refinement and prediction heads can realize the identity and the sigmoid gates may approach their endpoints, then $\mathcal{F}_H \cup \mathcal{F}_L \subseteq \overline{\mathcal{F}_{HL}}$. Consequently, for any nonnegative continuous risk $\mathcal{R}$,*

$$\inf_{f \in \mathcal{F}_{HL}} \mathcal{R}(f) \le \min\left\{ \inf_{f \in \mathcal{F}_H} \mathcal{R}(f), \inf_{f \in \mathcal{F}_L} \mathcal{R}(f) \right\}. \tag{4}$$

For an expressive classifier trained with log loss, the joint predictor strictly improves over the high-frequency branch if $I(Y_i; \ell_i \mid \widehat{h}_i) > 0$, and over the low-frequency branch if $I(Y_i; \widehat{h}_i \mid \ell_i) > 0$. Under Eq. equation 1, it is therefore strictly better than the best single branch.

3.2 Adaptive Hierarchical Global–Local Fusion

This stage introduces an Adaptive Hierarchical Global–Local Fusion (AHGF), which turns the two evidence streams into one node representation while allowing each node and feature dimension to select a different structural scale.

Frequency gate and residual injection. The first gate sees both the cross-conditioned local summary and the untouched low-frequency summary. After a residual MLP/LayerNorm block, the sampled-walk correction is injected through a second gate:

$$r_i^d = \mathrm{LN}\left(g_i^f \odot \widehat{h}_i + (1 - g_i^f) \odot \ell_i^0 + \mathrm{MLP}(g_i^f \odot \widehat{h}_i + (1 - g_i^f) \odot \ell_i^0)\right). \tag{5}$$

Where $\widetilde{r}_i^d = r_i^d + g_i^\Delta \odot \Delta_i$, $g_i^f = \sigma(W_f[\widehat{h}_i \| \ell_i^0] + b_f)$ and g_i^Δ is computed from $[r_i^d \| \widehat{h}_i \| \Delta_i]$. HDTE supplies candidate evidence, where AHGF decides how much of each candidate is trusted.

Self-token protection. Structural evidence must not erase target-specific attributes. We compare the gated structural representation with x_i and give the residual a separate gate $r_i^{\mathrm{sup}} = g_i^s \odot \widetilde{r}_i^d + (1 - g_i^s) \odot x_i + g_i^{\Delta s} \odot \Delta_i$, $g_i^s = \sigma(W_s[x_i \| \widetilde{r}_i^d] + b_s)$. An ambiguous neighborhood can therefore retain more of the self-token, and sampling noise cannot bypass self-token protection.

Agreement-aware Long-range topological injection. The Long-range topological token is queried against both frequency streams to obtain an aligned prior q_i. A learned gate and clipped cosine agreement then control injection: $z_i = r_i^{\mathrm{sup}} + \beta\, \omega_i\, g_i^g \odot q_i$, $\omega_i = \exp\left(\lambda_g \,\mathrm{clip}(\cos(q_i, r_i^{\mathrm{sup}}), -1, 1)\right)$. Here $g_i^g = \sigma(W_g[r_i^{\mathrm{sup}} \| q_i] + b_g)$ and $\beta = \sigma(\theta_g)$. The module flow is $(\widehat{h}_i, \ell_i^0, \Delta_i) \to r_i^d \to \widetilde{r}_i^d \to r_i^{\mathrm{sup}} \to z_i$; each arrow adds a new source only after the preceding representation has been gated. AHGF performs frequency selection, residual filtering, self-token preservation, and agreement-aware prior injection, providing the bounded interface between HDTE and metric constraints.

Theorem 2 (Bounded and stable adaptive fusion)**.** *Let $M_g(u, v) = g \odot u + (1 - g) \odot v$, with $g \in [0, 1]^d$, and assume the refinement map $T(u) = \mathrm{LN}(u + \mathrm{MLP}(u))$ satisfies $\|T(u)\|_2 \le L_T \|u\|_2 + c_T$ on the representation domain. For $\lambda_g, \kappa_s \ge 0$, the output in z_i satisfies*

$$\|z_i\|_2 \le L_T(\|\widehat{h}_i\|_2 + \|\ell_i^0\|_2) + c_T + \|x_i\|_2 + 2\|\Delta_i\|_2 + \beta e^{\lambda_g} \|q_i\|_2. \tag{6}$$

Moreover, $e^{-\lambda_g} \le \omega_i \le e^{\lambda_g}$ and ω_i is monotone in the cosine agreement. If the gate network is Lipschitz on a bounded domain, then M_g is also Lipschitz on that domain. Hence adaptive gating does not introduce an unbounded multiplicative amplification. A complete proof is given in Appendix A.3.

3.3 MULTISCALE PRIOR-INTERVAL CONSTRAINTS

This stage designs a Multiscale Prior-Interval Constraints (MPIC), which converts the forward representation into an explicitly shaped metric space: the fused node embedding is the anchor, the contextualized Long-range topological representation is the positive, and rank-dependent negatives define the separation interval.

Define metric roles. Let q_i be the aligned Long-range topological output from AHGF. The normalized anchor, positive, and pairwise distance are $a_i = \mathrm{norm}(P_a(\widetilde{r}_i^d))$, $p_i = \mathrm{norm}(q_i)$, $d_{ij} = \|a_i - p_j\|_2$.

The constraint is coupled to the forward workflow: fusion gates determine the anchor, while Long-range topological alignment determines the positive.

Fibonacci-ranked negatives and interval loss. Different-label candidates are sorted by increasing anchor–prior distance. A rank s_i is sampled uniformly from $\mathcal{F} = \{1, 2, 3, 5, 8, 13\}$ and wrapped modulo the candidate count when necessary. If π_i is the sorted index map, the selected negative and multiscale hinge are $r_i = 1 + ((s_i - 1) \bmod |\mathcal{N}_i|)$, $n_i = p_{\pi_i(r_i)}$

$$\mathcal{L}_{\mathrm{fib}} = \frac{1}{B} \sum_i \left[d(a_i, p_i) - d(a_i, n_i) + \frac{m_0}{\sqrt{s_i}} \right]_+ . \tag{7}$$

Where $\mathcal{N}_i = \{j : j \neq i, y_j \neq y_i\}$, with an all-nonself fallback if the set is empty. Smaller ranks impose larger margins, while larger ranks probe farther negatives.

Local preservation and joint optimization. For $c_{ik} = \cos(x_i, h_{ik})$, the local term rewards agreement between the self-token and each contextualized Hop-token. Combining this term with classification and warm-started prior alignment gives

$$\mathcal{L}_{\mathrm{loc}} = -\frac{1}{B} \sum_i \log \left[\frac{1}{K_h} \sum_k \exp(c_{ik}/\tau) + \epsilon \right] . \tag{8}$$

Thus, the optimization objective is as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{fib}}(t)\mathcal{L}_{\mathrm{fib}} + \lambda_{\mathrm{loc}}\mathcal{L}_{\mathrm{loc}}, \quad \lambda_{\mathrm{fib}}(t) = \alpha \min\left(1, \frac{t+1}{T_{\mathrm{warm}}}\right) \tag{9}$$

The implementation uses $\tau = 0.1$; $\mathcal{L}_{\mathrm{cls}}$ is the standard node-classification negative log-likelihood. The training flow is $z_i \to (a_i, p_i, n_i) \to (\mathcal{L}_{\mathrm{fib}}, \mathcal{L}_{\mathrm{loc}}) \to$ updated gates and encoders, closing the loop between representation construction and metric shaping. MPIC makes prior injection testable rather than implicit: Fibonacci-ranked negatives shape global intervals, the local term protects high-frequency semantics, and warm-up coordinates both signals so that the constraint refines rather than destabilizes the classifier.

Theorem 3 (Multiscale separation and local alignment)**.** *If $\mathcal{L}_{\mathrm{fib}} = 0$, every sampled triplet satisfies*

$$d(a_i, p_i) + \frac{m_0}{\sqrt{s_i}} \leq d(a_i, n_i). \tag{10}$$

For $s_i \in \mathcal{F}$, the enforced margin lies in $[m_0/\sqrt{13}, m_0]$ and its uniform-sampling expectation is $\frac{m_0}{|\mathcal{F}|} \sum_{s \in \mathcal{F}} s^{-1/2}$. Furthermore, each per-node term of Eq. equation 8 is bounded, and its partial derivative with respect to every c_{ik} is strictly negative. Minimizing it therefore increases self–Hop cosine agreement. See Appendix A.4 for the proof.

4 EXPERIMENT

We evaluate our proposed method on nine node-classification benchmarks. Six predominantly homophilous graphs are used for the main comparison, whereas Cornell, Wisconsin, and Actor test robustness under heterophily. results are test accuracy (%) reported as mean ± standard deviation over ten runs. Appendix B specifies the datasets, splits, metrics, implementation environment, and hyperparameters. Our code is available at `https://anonymous.4open.science/r/iclr2027-30F8`.

4.1 Comparison of Frequency Decoupling with Baseline

Table 1 compares the method in this paper with benchmark methods (detailed descriptions of the methods and datasets are in Appendix B.1) using node prediction tasks. Experiments for proposed method are presented as mean values and standard deviations run with 10 different random seeds. These results exceed most baseline methods with six datasets of varying sizes, proving the method's effectiveness.

Table 1: Comparison of our method with various baselines on various datasets. The best results are shown in **bold**. Given the lack of standard deviation values in the original papers (UGCformer, GUALformer, GT-SNT), part of the values for these are calculated by us. "-" and "X" mean unknown values.

Method	Year	Cora	Citeseer	Flickr	Photo	DBLP	Pubmed
GCN	2017	87.33±0.38	79.43±0.26	61.49±0.61	92.70±0.20	83.62±0.13	86.54±0.12
GAT	2017	86.29±0.53	80.13±0.62	54.29±2.56	93.87±0.11	84.19±0.19	86.32±0.16
GraphSAGE	2018	86.90±0.84	79.23±0.53	60.37±0.27	93.84±0.40	84.73±0.28	86.19±0.18
APPNP	2018	87.15±0.43	79.33±0.35	93.25±0.24	94.32±0.14	84.40±0.17	88.43±0.15
JKNet	2018	87.70±0.65	78.43±0.31	53.66±0.40	-	84.57±0.28	87.64±0.26
GPR-GNN	2021	88.27±0.40	78.46±0.88	-	94.49±0.14	-	89.34±0.25
GATv2	2022	-	-	-	-	84.96±0.47	85.75±0.55
GRAND+	2022	85.8x±0.4x	75.6x±0.4x	-	94.75±0.12	-	88.64±0.09
GNN-as-Judge	2026	79.54±0.39	74.39±1.63	-	-	-	87.49±1.23
NodePFN	2026	82.06±0.29	-	-	-	-	78.00±0.24
GT	2020	71.84±0.62	67.38±0.76	68.59±0.64	94.74±0.13	81.04±0.27	88.79±0.12
Graphormer	2021	72.85±0.76	66.21±0.83	66.16±0.24	92.74±0.14	80.93±0.39	82.76±0.24
SAN	2021	74.02±1.01	70.64±0.97	70.26±0.73	94.86±0.10	83.11±0.32	88.22±0.15
Gophormer	2021	87.85±0.10	80.23±0.09	91.51±0.28	-	85.20±0.20	89.40±0.14
ANS-GT	2022	88.60±0.45	80.25±0.39	-	94.41±0.62	-	89.56±0.55
GraphGPS	2022	-	-	-	95.06±0.13	-	88.94±0.16
Exphormer	2023	-	-	-	95.27±0.42	-	89.52±0.54
Gapformer	2023	87.37±0.76	76.21±1.47	-	94.81±0.45	**85.50±0.43**	88.98±0.46
NAGphormer	2023	90.56±0.39	80.02±0.80	89.66±0.63	95.49±0.11	84.62±0.13	89.60±0.14
UGCFormer	2025	84.94±0.43	73.41±0.27	51.58±0.36	96.21±0.31	76.95±0.57	81.79±0.81
DUALFormer	2025	85.88±0.10	74.45±0.39	51.42±0.49	**96.74±0.09**	83.57±0.11	83.97±0.43
GT-SNT	2026	84.7±0.8	74.0±0.5	50.02±0.17	94.35	84.47±0.27	80.6±0.4
Tokenphormer	2025	90.26±0.63	80.08±0.98	91.77±0.68	95.62±0.48	84.37±0.33	89.34±0.31
M3Dphormer	2025	88.48±1.94	77.53±1.56	52.73±0.86	95.91±0.68	84.32±0.49	89.96±0.4
Ours		**92.41±0.63**	**82.08±0.41**	**93.40±0.41**	96.27±0.22	85.15±0.17	**89.98±0.31**

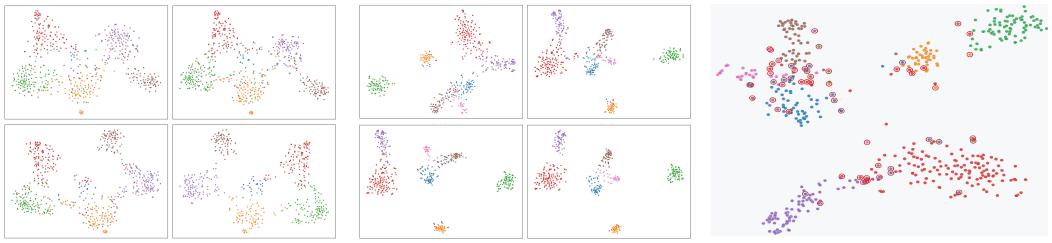
Table 1 compares the proposed method with message-passing GNNs and graph Transformer methods on six datasets of varying scales. Its performance outperforms most baseline methods, proving the model's effectiveness. Specifically, it achieves an average improvement of 1.31% over the baseline Tokenphormer across all six datasets and an average improvement of 9.02% over the latest neural network methods. The proposed method outperforms most GNN baseline methods due to two factors: (1) it enables extended graph topological capabilities by utilizing dual-frequency continuous representations; (2) it improves feature representation capabilities by enhancing feature fusion through adaptive attention.

The superiority of our approach over graph transformer baseline methods stems from the following advantages: (1) Our method utilizes dual-frequency decoupling to capture long-range topological information and graph structural information, enabling nodes to capture features with finer-grained characteristics. (2) Adaptive graph attention gating ensures consistent representation in the feature space. (3) Multiscale bounded constraints ensure alignment between local and global scales while preserving the flexibility to capture long-range topological information, thereby capturing complex details and correlations across the entire graph structure and achieving superior performance.

4.2 Ablation and Visualization Analysis

As shown in Table 5, our ablation experiments on a homogeneous dataset validated the effectiveness of our method design. The results indicate that the component methods in our approach complement one another, improving the model's performance by an average of 1.23% over the

variant baseline. Additionally, we compared four hybrid walking methods—uniform random walk, backtracking-free random walk, neighborhood-jumping walk, and backtracking-free neighborhood-jumping walk—with four single-walk methods. The results detailed in Appendix B.3, which demonstrate that removing any single component results in a decrease in accuracy, further prove that our methods do not over-rely on the Token.



(a) Citeseer (b) Cora (c) Cora without Walk-tokens

Figure 2: Representation analysis on test nodes. Each t-SNE panel reports a silhouette score and an intra- versus inter-class cosine-similarity gap, computed in the original embedding space before the two-dimensional projection. Circled nodes in (c) are misclassified.

Figure 2 visualizes node classification results on the Cora and Citeseer datasets. The classification boundaries become increasingly distinct as model components are added, further demonstrating the complementarity among the three components we propose. Moreover, integrating C from 2 reveals the following conclusions: 1) Our method addresses token over-reliance and optimizes the structure of long-token models; 2) It validates that the high- and low-frequency mechanisms can capture fine-grained structural and long-range topological information.

4.3 Token vs Accuracy

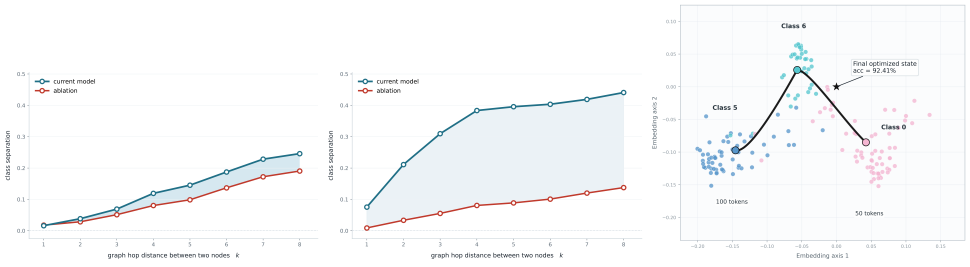
To compare with TokenPhormer while keeping other model settings constant as shown in Figure 4, results from all figures indicate that the method proposed in this paper achieves optimal results with the lowest token count on the majority datasets. Significant performance fluctuations were observed on a few datasets, which can be attributed to: 1) Our model does not over-rely on tokens. While increasing the walk budget expands the range of candidate low-frequency contexts, this may also introduce correlated or task-irrelevant walks; our residual attention gating makes the token budget irrelevant, thereby mitigating this effect. 2) Our model can learn long-range topological information from more tokens. While the Flickr dataset, which has strong contextual correlations, can achieve stability quickly, the Cora and DBLP datasets with numerous long-range dependencies show that the model demonstrates greater utility in leveraging extensive long-range topological information. A detailed analysis is provided in Appendix B.3.

4.4 Topological Expressiveness Comparison

The figure 3 shows that our model's ability to capture implicit and explicit topological information has been validated on the Cora and Citeseer datasets. As shown in a–b, our model can capture topological information between different nodes, where larger shaded areas between two line segments indicate clearer topological information. As shown in c, our model can capture topological information among homogeneous nodes, where increasingly concentrated classifications yield relatively clearer topological paths. Analysis of the experiments in provides further evidence of the model's effectiveness.

4.5 Heterogeneous Dataset Experiments

Table 2 shows our evaluation on heterogeneous datasets to determine whether the dual-frequency decoupling mechanism remains effective. Experimental results demonstrate that our method captures long-range topological information and graph structure, outperforming Tokenformer by an average of 1.82% on all three datasets. The standard deviation decreased from 2.30 to 1.64 on the Wisconsin dataset, indicating that our method does not assume any single branch to be globally reliable under heterogeneous conditions: it estimates the contribution of each node to the results,



(a) Citeseer (b) Cora (c) Cora class-centroid view.

Figure 3: Long-range structural analysis. Panels (a)–(b) compare the full model with an ablation that removes the long-range Walk branch; shaded bands show the plotted uncertainty. Panel (c) visualizes selected classes and their centroids in the final Cora embedding.

constraining the fused representation by context-specific priors. This further validates the rationality of our assumptions.

Table 2: Experiments with Heterogeneous Datasets

Method	Year	Cornell	Wisconsin	Actor
GCN	2017	45.67±7.96	52.55±4.27	28.73±1.17
GAT	2017	47.02±7.66	57.45±3.51	28.33±1.13
APPNP	2018	41.35±7.15	55.29±3.90	29.42±0.81
GATv2	2022	50.27±8.97	52.74±3.96	28.79±1.47
GAMMA	2025	80.79±0.64	86.67±0.89	36.09±0.94
NodePFN	2026	71.89±2.76	79.22±6.97	32.99±1.09
SAN	2021	50.85±8.54	51.37±3.08	27.12±2.59
Gapformer	2023	77.57±3.43	83.53±3.42	36.90±0.82
NAGphormer	2023	56.22±8.08	63.51±6.53	34.33±0.94
DUALFormer	2025	78.11±8.24	79.41±6.49	29.33±0.86
UGCFormer	2025	58.41±3.28	61.28±5.87	30.63±0.62
GT-SNT	2026	78.78±4.95	77.25±6.35	28.74±1.10
Tokenphormer	2025	76.22±2.13	86.17±2.30	37.01±0.83
Ours		**81.08±1.24**	**86.27±1.64**	**37.50±0.84**

4.6 Conclusion

We propose a dual-frequency decoupling adaptive constraint method, which overcomes the challenge of inaccurate representations resulting from the task-irrelevance of long-range topological information in graph transformer approaches. By decoupling high- and low-frequency graph structures and topological information, it ensures fine-grained extraction of feature representations. Residual attention gating maintains feature consistency, improving task representation capabilities. Multi-scale consistency ensures alignment and training optimization, guaranteeing the model's stability.

Furthermore, since this method was primarily evaluated on node classification tasks, its performance in inductive updating, dynamic graphs, and larger-scale streaming scenarios remains to be determined. Future work will explore learnable frequency definitions, more efficient long-range token selection, and extensions to temporal, inductive, and graph-level prediction tasks.

AI use statement

In this work, we used generative AI tools solely to improve the readability of the manuscript and to format its references. We did not use generative AI tools to discover research topics, to identify research gaps, or to draft any part of the paper; these uses are not applicable to this work. All AI-assisted content was carefully reviewed, word by word, to ensure that it faithfully preserves the

original meaning of our research. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

Reproducibility statement

This paper provides a link to anonymized code (see Section 4 for details), and the parameter lists and hardware/software environments are available in the appendix. Using the materials provided above, all experimental procedures and results in this paper can be fully reproduced.

References

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*, 2021.

Deyu Bo, Xiao Wang, Chuan Shi, and Huawei Shen. Beyond low-frequency information in graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 3950–3957, 2021.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022.

Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3438–3445, 2020.

Jinsong Chen, Kaiyuan Gao, Gaichao Li, and Kun He. NAGphormer: A tokenized graph transformer for node classification in large graphs. In *International Conference on Learning Representations*, 2023.

Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized pagerank graph neural network. In *International Conference on Learning Representations*, 2021.

Jeongwhan Choi, Jongwoo Kim, Woosung Kang, and Noseong Park. Learning posterior predictive distributions for node classification from synthetic graph priors. In *The Fourteenth International Conference on Learning Representations*, 2026.

Chenhui Deng, Zichao Yue, and Zhiru Zhang. Polynormer: Polynomial-expressive graph transformer in linear time. In *International Conference on Learning Representations*, 2024.

Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. In *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*, 2021.

Wenzheng Feng, Yuxiao Dong, Tinglin Huang, Ziqi Yin, Xu Cheng, Evgeny Kharlamov, and Jie Tang. Grand+: Scalable graph random neural networks. In *Proceedings of the ACM web conference 2022*, pp. 3248–3258, 2022.

Dongqi Fu, Zhigang Hua, Yan Xie, Jin Fang, Si Zhang, Kaan Sancak, Hao Wu, Andrey Malevich, Jingrui He, and Bo Long. Vcr-graphormer: A mini-batch graph transformer via virtual connections. In *International Conference on Learning Representations*, 2024.

Amir Ghazizadeh, Rickard Ewetz, and Hao Zheng. Gamma: Gated multi-hop message passing for homophily-agnostic node representation in gnns. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1263–1272, 2017.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Siyuan Huang, Yunchong Song, Jiayue Zhou, and Zhouhan Lin. Cluster-wise graph transformer with dual-granularity kernelized attention. In *Advances in Neural Information Processing Systems*, volume 37, pp. 33376–33401, 2024.

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

Johannes Klicpera, Aleksandar Bojchevski, and Stephan Günnemann. Combining neural networks with personalized pagerank for classification on graphs.

Devin Kreuzer, Dominique Beaini, William L. Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking graph transformers with spectral attention. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21618–21629, 2021.

Chuang Liu, Yibing Zhan, Xueqi Ma, Liang Ding, Dapeng Tao, Jia Wu, and Wenbin Hu. Gapformer: graph transformer with graph pooling for node classification. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 2196–2205. International Joint Conferences on Artificial Intelligence, 2023. 32nd International Joint Conference on Artificial Intelligence, IJCAI 2023 ; Conference date: 19-08-2023 Through 25-08-2023.

Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *Advances in neural information processing systems*, 35:14501–14515, 2022.

Hamed Shirzad, Ameya Velingker, Balaji Venkatachalam, Danica J. Sutherland, and Ali Kemal Sinop. Exphormer: Sparse transformers for graphs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31613–31632. PMLR, 2023.

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

Yujie Xing, Xiao Wang, Yibo Li, Hai Huang, and Chuan Shi. Less is more: On the over-globalizing problem in graph transformers. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 54656–54672, 2024.

Yujie Xing, Xiao Wang, Bin Wu, Hai Huang, and Chuan Shi. Unifying and enhancing graph transformers via a hierarchical mask framework. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *International conference on machine learning*, pp. 5453–5462. pmlr, 2018.

Ruiyao Xu and Kaize Ding. GNN-as-judge: Unleashing the power of LLMs for graph learning with GNN feedback. In *The Fourteenth International Conference on Learning Representations*, 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? In *Advances in Neural Information Processing Systems*, volume 34, pp. 28877–28888, 2021.

Huizhe Zhang, Jintang Li, Yuchang Zhu, Huazhen Zhong, and Liang Chen. Gt-snt: A linear-time transformer for large-scale graphs via spiking node tokenization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 16298–16306, 2026.

Zaixi Zhang, Qi Liu, Qingyong Hu, and Chee-Kong Lee. Hierarchical graph transformer with adaptive node sampling. *Advances in Neural Information Processing Systems*, 35:21171–21183, 2022.

Jianan Zhao, Chaozhuo Li, Qianlong Wen, Yiqi Wang, Yuming Liu, Hao Sun, Xing Xie, and Yanfang Ye. Gophormer: Ego-graph transformer for node classification. *arXiv preprint arXiv:2110.13094*, 2021.

Zijie Zhou, Zhaoqi Lu, Xuekai Wei, Rongqin Chen, Shenghui Zhang, Pak Lon Ip, and Leong Hou U. Tokenphormer: Structure-aware multi-token graph transformer for node classification. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, volume 39, pp. 13428–13436, 2025.

Jiaming Zhuo, Yuwei Liu, Yintong Lu, Ziyi Ma, Kun Fu, Chuan Wang, Yuanfang Guo, Zhen Wang, Xiaochun Cao, and Liang Yang. Dualformer: Dual graph transformer. In *International Conference on Learning Representations*, 2025a.

Jiaming Zhuo, Ziyi Ma, Yintong Lu, Yuwei Liu, Kun Fu, Di Jin, Chuan Wang, Wu Wenning, Zhen Wang, Xiaochun Cao, and Liang Yang. A closer look at graph transformers: Cross-aggregation and beyond. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

A Theoretical Analysis

This appendix proves the three results stated in Section 3. The results characterize representational capacity and boundedness; they do not assume that a nonconvex optimizer reaches a global optimum.

The revised workflow uses the following interfaces: HDTE outputs $\widehat{h}_i$ and ℓ_i^0, AHGF maps them to z_i through the gated residual path, and MPIC treats the resulting representation as the anchor for prior alignment. The proofs below use these interfaces only; standard attention, pooling, and classification operations are absorbed into the named maps in the main text.

A.1 Justification of the Two Workflow Assumptions

The two assumptions in Section 3 are modeling conditions rather than claims that hold for every graph. This appendix establishes their statistical meaning and proves the boundedness consequences required by the architecture.

Assumption 1: conditional complementarity. Let Y be the node label and H, L the high- and low-frequency summaries. We assume $I(Y; L \mid H) > 0$ and $I(Y; H \mid L) > 0$ on a set of positive probability. These inequalities are exactly the conditions for strict reductions in optimal log-loss risk when the missing view is added.

Proposition 1. Under log loss, the Bayes risks based on H, L, and (H, L) are $H(Y \mid H)$, $H(Y \mid L)$, and $H(Y \mid H, L)$. The chain rule gives

$$H(Y \mid H, L) = H(Y \mid H) - I(Y; L \mid H) = H(Y \mid L) - I(Y; H \mid L). \tag{11}$$

Proof. The identity follows from the chain rule for conditional entropy. Each conditional mutual information is an expected Kullback–Leibler divergence and is therefore nonnegative. The joint Bayes risk is strictly smaller than the high-frequency risk exactly when $I(Y; L \mid H) > 0$, and strictly smaller than the low-frequency risk exactly when $I(Y; H \mid L) > 0$. Hence Assumption 1 is the minimal information condition used by Theorem 1. □

The condition is empirically auditable: held-out probes can estimate the three conditional risks or their accuracy analogues, and positive reductions provide evidence for complementarity without changing the predictor.

Assumption 2: bounded gated prior fusion. Inputs and residuals are finite, gate networks are locally Lipschitz, and adaptive coefficients are sigmoid outputs. The Long-range topological prior may be imperfect; only clipped cosine agreement controls its amplitude.

Proposition 2. With $c_i = clip(cos(q_i, r_i^{\mathrm{sup}}), -1, 1)$, the agreement factor in ω_i satisfies

$$e^{-\lambda_g} \le \omega_i = e^{\lambda_g c_i} \le e^{\lambda_g}. \tag{12}$$

Proof. Sigmoid outputs lie in $(0, 1)$, so gated mixtures have operator norm at most one and the triangle inequality gives $||g \odot u + (1-g) \odot v||_2 \leq ||u||_2 + ||v||_2$. Cosine clipping places c_i in $[-1, 1]$, which proves the displayed exponential bounds. Finite linear maps, tanh, sigmoid, clipping, cosine on nonzero vectors, and LayerNorm with a positive stabilizer are locally Lipschitz on bounded domains; their finite composition is locally Lipschitz. The complete residual is therefore bounded by Eq. equation 6. □

The imperfect-teacher clause is structurally justified because MPIC does not force the prior to equal the anchor: it only requires a rank-dependent positive-negative interval, while the agreement factor attenuates incompatible priors. Thus Assumption 2 is sufficient for stability while allowing noisy Long-range topological components.

A.2 Proof of Theorem 1

Proof. We first show that the joint architecture contains either single-frequency predictor in its closure. Consider a predictor based only on the high-frequency representation. Set the cross-frequency attention output to zero, let the frequency gate approach the all-one vector, let the refinement map realize the identity, and suppress the self-token, sampling residual, and Long-range topological residual through their corresponding gates. A sigmoid gate approaches the all-one vector as its bias tends to $+\infty$. The remaining prediction head therefore realizes the high-frequency predictor to arbitrary precision. The same construction with the frequency-gate bias tending to $-\infty$ recovers a low-frequency predictor. Hence

$$\mathcal{F}_H \cup \mathcal{F}_L \subseteq \overline{\mathcal{F}_{HL}}. \tag{13}$$

For a continuous risk, taking the infimum over a class cannot give a larger value than taking the infimum over either of its subclasses. Applying this observation to Eq. equation 13 proves Eq. equation 4.

It remains to characterize when the inequality can be strict. With log loss and an expressive prediction head, the Bayes risks based on the two individual views and their joint representation are, respectively,

$$H(Y_i \mid \widehat{h}_i), \qquad H(Y_i \mid \ell_i), \qquad H(Y_i \mid \widehat{h}_i, \ell_i). \tag{14}$$

The chain rule for conditional mutual information gives

$$H(Y_i \mid \widehat{h}_i, \ell_i) = H(Y_i \mid \widehat{h}_i) - I(Y_i; \ell_i \mid \widehat{h}_i), \tag{15}$$

$$= H(Y_i \mid \ell_i) - I(Y_i; \widehat{h}_i \mid \ell_i). \tag{16}$$

Conditional mutual information is nonnegative, so the joint Bayes risk is no larger than either single-view risk. If the corresponding term in Eq. equation 16 is positive, the reduction relative to that branch is strict. Under Eq. equation 1, it is strict relative to both branches. This proves the claim. □

Remark. The theorem is a capacity statement. It establishes that joint encoding need not sacrifice either branch and quantifies the possible Bayes-risk gain; it does not claim that every parameter setting preserves all information or that gradient descent necessarily finds the optimal setting.

A.3 Proof of Theorem 2

Proof. For any $u, v \in \mathbb{R}^d$ and $g \in [0, 1]^d$, the triangle inequality and the fact that the diagonal operators induced by g and $1 - g$ have spectral norm at most one yield

$$\|M_g(u, v)\|_2 \leq \|g \odot u\|_2 + \|(1 - g) \odot v\|_2 \leq \|u\|_2 + \|v\|_2. \tag{17}$$

Equation equation 17 applied to the frequency gate gives

$$\|r_i^f\|_2 \leq \|\widehat{h}_i\|_2 + \|\ell_i^0\|_2. \tag{18}$$

The assumed refinement bound then implies

$$\|r_i^d\|_2 \leq L_T(\|\widehat{h}_i\|_2 + \|\ell_i^0\|_2) + c_T. \tag{19}$$

Because every sampling gate is also coordinate-wise bounded by one, the residual and self-protection gates in the workflow (Eq. equation 5 and $r_i^{\sup}$) give

$$\|\widetilde{r}_i^d\|_2 \le \|r_i^d\|_2 + \|\Delta_i\|_2, \tag{20}$$

$$\|r_i^{\sup}\|_2 \le \|\widetilde{r}_i^d\|_2 + \|x_i\|_2 + \|\Delta_i\|_2. \tag{21}$$

For completeness, Eq. equation 3 also satisfies

$$\|\Delta_i\|_2 \le \rho_s e^{\kappa_s}(\|\bar{\ell}_i\|_2 + \|\ell_i^0\|_2), \tag{22}$$

because cosine similarity lies in $[-1, 1]$.

Let $c_i = \mathrm{clip}(\cos(q_i, r_i^{\sup}), -1, 1)$. Since $c_i \in [-1, 1]$ and $\lambda_g \ge 0$,

$$e^{-\lambda_g} \le \omega_i = e^{\lambda_g c_i} \le e^{\lambda_g}, \qquad \frac{\partial \omega_i}{\partial c_i} = \lambda_g e^{\lambda_g c_i} \ge 0. \tag{23}$$

The Long-range topological gate has operator norm at most one and $0 < \beta < 1$, hence

$$\|z_i - r_i^{\sup}\|_2 = \|\beta \omega_i g_i^g \odot q_i\|_2 \le \beta e^{\lambda_g} \|q_i\|_2. \tag{24}$$

Combining Eqs. equation 19, equation 21, and equation 24 gives Eq. equation 6.

We finally make the stability statement explicit. Suppose $\|u\|_2, \|v\|_2, \|u'\|_2, \|v'\|_2 \le R$, and the input-dependent gate is L_g-Lipschitz in the concatenated Euclidean norm. Writing $g = g(u, v)$ and $g' = g(u', v')$,

$$\|M_g(u, v) - M_{g'}(u', v')\|_2 \tag{25}$$

$$\le \|u - u'\|_2 + \|v - v'\|_2 + \|g - g'\|_\infty \|u' - v'\|_2 \tag{26}$$

$$\le (\sqrt{2} + 2RL_g)\|[u - u' \| v - v']\|_2. \tag{27}$$

Thus the adaptive mixture is Lipschitz on a bounded representation domain. Linear maps, sigmoid, attention with finite parameters, and LayerNorm with a positive numerical stabilizer are locally Lipschitz; their finite composition therefore preserves this property. This completes the proof. □

Why no convergence claim is made. The bounds above establish controlled feature magnitudes and perturbation stability. They do not establish convergence of stochastic optimization, which would require separate assumptions on the objective, gradient noise, and learning-rate schedule.

A.4 Proof of Theorem 3

Proof. For every sampled triplet, define

$$\xi_i = \left[d(a_i, p_i) - d(a_i, n_i) + m_0/\sqrt{s_i}\right]_+ . \tag{28}$$

Each ξ_i is nonnegative. Therefore, $\mathcal{L}_{\mathrm{fib}} = B^{-1} \sum_i \xi_i = 0$ implies $\xi_i = 0$ for every sampled triplet. By the definition of the hinge,

$$d(a_i, p_i) - d(a_i, n_i) + \frac{m_0}{\sqrt{s_i}} \le 0, \tag{29}$$

which rearranges to Eq. equation 10.

The function $m(s) = m_0 s^{-1/2}$ is decreasing for $s > 0$. Since the smallest and largest elements of $\mathcal{F}$ are 1 and 13,

$$\frac{m_0}{\sqrt{13}} \le m(s_i) \le m_0. \tag{30}$$

Uniform sampling from $\mathcal{F}$ gives

$$\mathbb{E}[m(s_i)] = \frac{m_0}{|\mathcal{F}|} \sum_{s \in \mathcal{F}} s^{-1/2}. \tag{31}$$

When $|\mathcal{N}_i| \ge 13$, Eq. equation 7 reduces to $r_i = s_i$, so sorted distance and margin are coupled monotonically: closer negatives receive larger margins. If $|\mathcal{N}_i| < 13$, modulo wrapping retains

the interval in Eq. equation 30 but not this rank monotonicity. This is the exact behavior of the implementation.

It remains to prove the local-alignment statement. Let

$$A_i = \frac{1}{K_h} \sum_{k=1}^{K_h} \exp(c_{ik}/\tau) + \epsilon, \qquad \ell_i^{\mathrm{loc}} = -\log A_i. \tag{32}$$

Cosine similarity satisfies $c_{ik} \in [-1, 1]$. Hence

$$e^{-1/\tau} + \epsilon \le A_i \le e^{1/\tau} + \epsilon \tag{33}$$

and therefore

$$-\log(e^{1/\tau} + \epsilon) \le \ell_i^{\mathrm{loc}} \le -\log(e^{-1/\tau} + \epsilon). \tag{34}$$

Thus the local term is finite and bounded. Direct differentiation gives

$$\frac{\partial \ell_i^{\mathrm{loc}}}{\partial c_{ik}} = -\frac{\exp(c_{ik}/\tau)}{\tau K_h A_i} < 0. \tag{35}$$

Consequently, decreasing the objective pushes each self–Hop cosine similarity upward, subject to the other terms in the joint objective. This proves the local-alignment claim and completes the theorem. □

B Experimental Details

B.1 Experiment Setup

Datasets We evaluate nine node-classification graphs covering different scales, feature dimensions, class counts, and structural regimes. Cora and Citeseer are small citation networks with sparse bag-of-words features; Pubmed is a larger citation graph with three classes. Amazon Photo is a dense product-co-purchase graph, Flickr is a social-image graph with high-dimensional features, and DBLP is a larger author collaboration graph. These six graphs are predominantly homogeneous and form the main comparison in Table 1. Cornell and Wisconsin are small heterogeneous WebKB graphs, while Actor is a larger actor-co-occurrence graph with weak neighborhood-label agreement. They form the robustness comparison in Table 2. Node counts, edge entries, feature dimensions, class counts, and regime labels are reported in Table 3.

Table 3: Datasets used in the homogeneous and Heterogeneous benchmark comparisons. Edge entries are directed adjacency entries after graph preprocessing; input dimensions include the appended three-dimensional positional encoding where applicable. The split column reports train/validation/test node counts, with official split 0 shown for the Heterogeneous datasets.

Dataset	Nodes	Edge entries	Input dim.	Classes	Regime	Train/Val/Test
Cora	2,708	10,556	1,436	7	Homogeneous	1,626/542/540
Citeseer	3,327	9,228	3,706	6	Homogeneous	1,997/666/664
Pubmed	19,717	88,651	503	3	Homogeneous	11,830/3,944/3,943
Photo	7,650	238,163	748	8	Homogeneous	4,591/1,530/1,529
Flickr	7,575	479,476	12,047	9	Homogeneous	4,545/1,515/1,515
DBLP	17,716	105,734	1,639	4	Homogeneous	10,629/3,543/3,544
Cornell	183	554	1,706	5	Heterogeneous	87/59/37
Wisconsin	251	900	1,706	5	Heterogeneous	120/80/51
Actor	7,600	53,318	935	5	Heterogeneous	3,648/2,432/1,520

Baseline The comparison methods in Tables 1 and 2 are organized according to how they expose graph structure to the predictor. Results copied from prior work retain the method names and reported availability in the tables; a dash indicates that the corresponding result was not available in the source comparison.

Message-passing GNNs. GCN, GAT, and GraphSAGE represent the standard neighborhood-aggregation family, using spectral propagation, learned neighbor attention, and sampled inductive

aggregation, respectively (Kipf & Welling, 2017; Veličković et al., 2018; Hamilton et al., 2017). APPNP (Klicpera et al.), JKNet (Xu et al., 2018), GPR-GNN (Chien et al., 2021), GATv2 (Brody et al., 2022), GNN-as-Judge (Xu & Ding, 2026) , NodePFN (Choi et al., 2026), and GRAND+ (Feng et al., 2022) extend this family with personalized propagation, layer-wise aggregation, adaptive propagation coefficients, dynamic attention, or stochastic augmentation. These baselines provide the local receptive-field reference for both the homogeneous and heterogeneous tables: they test whether our method's gains remain after a conventional message-passing encoder has already extracted neighborhood evidence.

Graph Transformers. GT, Graphormer, SAN (Dwivedi & Bresson, 2021; Ying et al., 2021; Kreuzer et al., 2021), and GT-SNT (Zhang et al., 2026) use attention to enlarge the receptive field, with graph-specific positional or structural encodings distinguishing them from sequence Transformers. ANS-GT (Zhang et al., 2022), GraphGPS (Rampášek et al., 2022), Exphormer (Shirzad et al., 2023), and Gapformer (Liu et al., 2023) further vary the attention pattern or sparsify global interactions; Exphormer and Gaphormer are representative examples of structured or virtual connectivity for scalable global context. These methods test the long-range alternative to local message passing, but they do not all expose a separate local and long-range stream. The heterogeneous table additionally includes GAMMA (Ghazizadeh et al., 2025), a recent graph-Transformer baseline designed for challenging neighborhood-label relationships.

Token-based and multi-scale Graph Transformers. Gophormer (Zhao et al., 2021), NAGphormer (Chen et al., 2023), Tokenphormer, M3Dphormer (Xing et al., 2025), UGCFormer (Zhuo et al., 2025b), and DUALFormer (Zhuo et al., 2025a) construct node-centered structural tokens or explicitly combine multiple interaction scales. NAGphormer uses hop-wise tokens, whereas Tokenphormer combines several structure-aware token sources; these designs motivate the controlled comparison with ours because they expose a stronger tokenized backbone than a single global attention matrix (Chen et al., 2023; Zhou et al., 2025). Related recent work explores polynomially expressive attention, cluster-level interactions, and hierarchical masks (Deng et al., 2024; Huang et al., 2024; Xing et al., 2025). The remaining tokenized rows in the two tables are included as reported recent baselines so that the comparison covers both homogeneous and heterogeneous regimes. our method differs in the point of fusion: it first keeps Hop- and Walk-derived evidence in separate frequency streams, then performs node-wise adaptive fusion and prior-constrained alignment rather than averaging all token sources in one undifferentiated pool.

Benchmark Metrics The metrics below are matched to the experiment that they diagnose. Accuracy is the task metric for the baseline comparisons in Tables 1 and 2, the component ablations in Table 5, and the Walk-token budget sweep in Figure 4. Geometry metrics are used in the branch-representation analysis in Figure 2, whereas distance-conditioned separation is used in Figure 3 to isolate long-range structural expressiveness.

Classification accuracy. For a transductive node-classification test set, we report

$$\mathrm{Acc} = \frac{1}{|\mathcal{V}_{\mathrm{test}}|} \sum_{i \in \mathcal{V}_{\mathrm{test}}} \mathbb{I}\Big[\arg\max_c p_{ic} = y_i\Big]. \tag{36}$$

All accuracy values in the paper are percentages. In Figure 4, the Walk-token budget K_w is the horizontal variable and Eq. equation 36 is the vertical metric; all other hyperparameters are fixed within each curve. The dashed portion of the NAGphormer curve marks budgets beyond the available exact-hop range, rather than additional distinct hop information.

Representation geometry: cosine gap and silhouette. Given test embeddings $\{\mathbf{z}_i\}$ and labels $\{y_i\}$, define

$$\mathcal{P}_{\mathrm{intra}} = \{(i,j) : y_i = y_j,\ i \neq j\}, \qquad S_{\mathrm{intra}} = \frac{1}{|\mathcal{P}_{\mathrm{intra}}|} \sum_{(i,j)\in\mathcal{P}_{\mathrm{intra}}} \cos(\mathbf{z}_i, \mathbf{z}_j), \tag{37}$$

$$\mathcal{P}_{\mathrm{inter}} = \{(i,j) : y_i \neq y_j\}, \qquad S_{\mathrm{inter}} = \frac{1}{|\mathcal{P}_{\mathrm{inter}}|} \sum_{(i,j)\in\mathcal{P}_{\mathrm{inter}}} \cos(\mathbf{z}_i, \mathbf{z}_j). \tag{38}$$

The cosine gap is

$$\mathrm{Gap}_{\cos} = S_{\mathrm{intra}} - S_{\mathrm{inter}}. \tag{39}$$

A larger positive gap indicates that same-class pairs are more aligned than different-class pairs. Self-pairs are excluded. We also ℓ_2-normalize the original high-dimensional embeddings and use cosine distance. For node i, let $a(i)$ be its mean distance to other nodes of the same class and let $b(i)$ be the minimum, over all other classes, of its mean distance to that class. We report

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \qquad \mathrm{Sil} = \frac{1}{N}\sum_{i=1}^{N} s(i). \tag{40}$$

The score lies in $[-1, 1]$, and larger is better. The t-SNE coordinates in Figure 2 are used only for visualization: both Eq. equation 39 and Eq. equation 40 are evaluated before the two-dimensional projection, thereby avoiding projection-induced changes.

Long-range topology: distance-conditioned class separation. Let $\mathcal{P}_k = \{(i, j) : d_G(i, j) = k\}$ be node pairs at exact shortest-path distance k. Splitting this set into same- and different-class pairs gives

$$\mathrm{Sep}(k) = \frac{1}{|\mathcal{P}_k^+|}\sum_{(i,j)\in\mathcal{P}_k^+} \cos(\mathbf{z}_i, \mathbf{z}_j) - \frac{1}{|\mathcal{P}_k^-|}\sum_{(i,j)\in\mathcal{P}_k^-} \cos(\mathbf{z}_i, \mathbf{z}_j). \tag{41}$$

This is the metric in Figures 3a–b. It measures whether class-compatible geometry is retained at each topological scale; a positive and increasing value indicates separation rather than undifferentiated smoothing. In particular, it tests the long-range integrity question addressed by the Walk branch, not merely the final classification boundary.

B.2 Experiment Details

Data processing and reporting. The loader converts every graph to a bidirected representation before token generation, so "edge entries" in Table 3 counts directed adjacency entries rather than unique undirected edges. Features are row-normalized. A three-dimensional Laplacian positional encoding is appended on Cora, Citeseer, Pubmed, Photo, Cornell, Wisconsin, and Actor; Flickr and DBLP use their loaded feature matrices without this appended encoding. The six homogeneous datasets use stratified 60/20/20 train/validation/test splits, whereas Cornell, Wisconsin, and Actor use the ten official Geom-GCN splits. Self-, Hop-, Walk-, and SGPM-tokens are generated from the same processed graph within each run. We select the checkpoint by validation accuracy, evaluate the held-out test nodes once, and report the mean and standard deviation over ten runs. The token-budget curves report the accuracy at each displayed budget and are not confidence intervals.

We fix the model and optimization settings before each comparison so that the reported differences are attributable to the structural modules. The implementation uses Python 3.7.12, PyTorch 1.10.1 with CUDA 11.3, DGL 0.9.0, PyTorch Geometric 2.0.4, NumPy 1.21.5, SciPy 1.7.3, scikit-learn 1.0.2, pandas 1.3.5, NetworkX 2.6.3, and Matplotlib 3.5.1. Experiments were executed on 64-bit Ubuntu 24.04 with an NVIDIA GeForce RTX 3090 (24 GB). The host driver reports CUDA 13.0 compatibility, whereas the isolated training environment uses the CUDA 11.3 toolkit; the latter is the relevant runtime version for PyTorch.

All reported configurations use a one-layer Transformer with hidden dimension 512 and SGPM-tokens. Optimization uses AdamW, a linear polynomial-decay schedule with 400 warm-up updates and 1,000 total updates, an end learning rate of 10^{-4}, at most 2,000 epochs, and early stopping with patience 50. MPIC uses base margin $m_0 = 1$, Fibonacci ranks $\{1, 2, 3, 5, 8, 13\}$, local temperature $\tau = 0.1$, and local-loss weight $\lambda_{\mathrm{loc}} = 0.1$. The triplet coefficient is linearly warmed from zero to the dataset-specific α. Mixed Walk-tokens allocate the stated proportions to uniform random walk (URW), non-backtracking random walk (NBRW), and neighborhood-jump walk (NJW); the remainder is assigned to non-backtracking neighborhood-jump walk (NBNJW).

Table 4 records the commands used for the six principal homogeneous datasets. The random-walk cache itself is generated with seed 42; the reported run seed controls model initialization and mini-batch sampling. In the sensitivity experiment, only K_w is replaced by the budget on the horizontal axis.

Table 4: Dataset-specific implementation settings.

	Homogeneous datasets						Heterogeneous datasets		
Dataset	**Cora**	**Citeseer**	**Pubmed**	**Flickr**	**Photo**	**DBLP**	**Cornell**	**Actor**	**Wisconsin**
Test Acc	0.9241	0.8208	0.8998	0.93399	0.9627	0.8515	0.8108	0.3750	0.8627
t_nums	40	100	100	30	110	20	100	15	100
w_len	4	4	4	2	5	3	4	2	4
hop_num	3	3	3	2	3	3	3	2	3
seed	5	0	42	0	0	0	1	2	9
peak_lr	0.005	0.005	0.002	0.0009	0.001	0.002	0.005	0.001	0.005
weight_decay	1e-5	1e-5	5e-4	5e-4	5e-4	5e-4	1e-5	5e-4	1e-5
n_heads	2	1	2	4	4	2	1	4	1
α	0.1	0.1	0.05	0.04	0.05	0.05	0.01	0.01	0.1
triplet_warmup_epochs	30	30	50	20	50	50	10	10	30
batch_size	2000	2000	1000	512	1000	1000	512	512	512
dropout / attn_dropout	–	–	–	–	–	–	–	0.5 / 0.5	–
sgpm_token	True	True	True	True	True	True	True	True	True/False
use_sampling_branch	False	False	False	True	True	True	False	False	–

B.3 Supplementary Experimental Details

This section records the controlled experiments that diagnose the three scientific questions raised in the main text: whether local and long-range signals are complementary, whether adaptive gating is needed to combine them, and how much long-range token coverage is useful before irrelevant context enters the representation. All variants use the same splits, optimizer, and checkpoint rule as the full model. The first block removes mechanisms from the shared backbone, while the second block keeps the full model and changes only the Walk sampler.

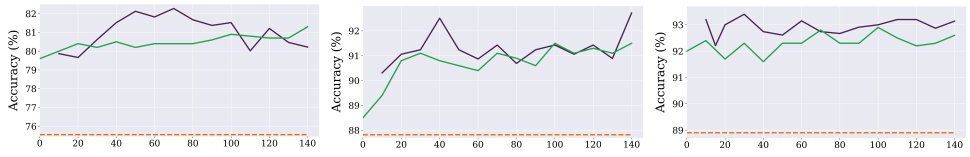
Table 5 shows that removing HDTE, AHGF, and MPIC together reduces the mean accuracy from 88.16 to 86.71%, a 1.44-point drop. Keeping only HDTE, AHGF, or MPIC recovers 0.82, 0.54, or 0.98 points over this common backbone, respectively, while the full model improves over those isolated variants by 0.63, 0.90, and 0.47 points. This pattern supports a complementary workflow: HDTE preserves frequency-specific evidence, AHGF decides which stream is useful for each node, and MPIC prevents the fusion from violating the local and long-range consistency constraints. The single-walk rows address the coverage question directly. NJW is the strongest single policy on all three datasets (80.92/90.80/89.50), but the mixed sampler improves over NJW by 1.16, 1.61, and 0.48 points, indicating that distinct random-walk policies supply complementary long-range routes.

Table 5: Ablation experiment.

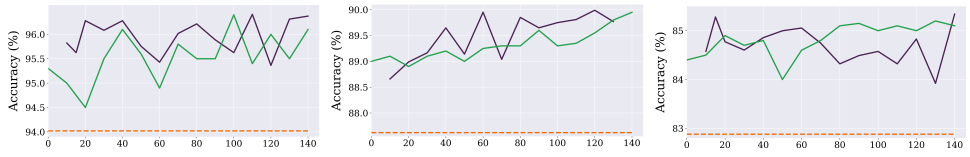
	Citeseer	cora	pumbed	Flickr	Photo	DBLP
Ours	82.08	92.41	89.98	93.40	96.27	85.15
$w/o\ AHGF + MPIC$	81.48	91.48	89.63	92.2	95.75	84.4
$w/o\ HDTE + MPIC$	80.57	91.67	89.53	93.2	95.77	84.6
$w/o\ AHGF + MPIC$	81.33	92.04	89.70	92.40	96.14	84.4
$w/o\ all$	80.33	90.74	89.07	91.77	95.62	84.37
only URW	80.69	90.41	89.33	92.07	95.83	84.90
only NBRW	80.03	89.72	89.26	92.27	95.74	84.86
only NJW	80.92	90.80	89.50	92.38	95.61	84.70
NBNJW	80.56	89.63	89.36	92.09	96.00	84.71

The token-budget experiment in Figure 4 holds the encoder, loss, and sampler proportions fixed and varies only K_w. The best displayed our values are 82.28 on Citeseer with 70 tokens, 92.72 on Cora with 140, 93.40 on Flickr with 30, 96.41 on Photo with 110, 89.99 on Pubmed with 120, and 85.34 on DBLP with 140. The peak-to-trough ranges are 2.61, 2.41, 1.19, 1.05, 1.33, and 1.42 points, respectively. Flickr therefore reaches its best result with a compact context, whereas Cora and DBLP benefit from a broader candidate pool. The curves are not monotone: increasing K_w first improves low-frequency coverage and then admits correlated or task-irrelevant walks. This behavior links the

sweep to the framework's scientific question: long-range integrity requires adaptive refinement and gating, not an unrestricted increase in token count.



(a) Citeseer (b) cora (c) flickr



(d) Photo (e) pubmed (f) DBLP

Figure 4: Test accuracy versus Walk-token budget. Purple, green, and orange denote our method, Tokenphormer, and NAGphormer, respectively. Dashed orange segments indicate budgets beyond the available hop-token range. Comparisons are made from the displayed curves; Appendix B.1 defines the sweep protocol.