TOWARD OMNI MULTIMODAL GRAPH FOUNDATION MODEL: A TOPOLOGY-DRIVEN BINDING APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal graph foundation models (MGFMs) seek to learn generalizable representations from large-scale graphs with heterogeneous node modalities. However, real-world Multimodal-Attributed Graphs (MAGs) often contain incomplete node attributes, limiting the scale and diversity of available pretraining corpora. Besides, existing MGFMs primarily incorporate graph topology as structural context, overlooking its role in guiding multimodal binding and shaping a unified representation space. To address these challenges, we propose **GraphBind**, a topology-driven approach that uses graph topology to bind rich modality information into a unified shared space. GraphBind is motivated by the stability of graph topology, which provides structural references and complementary semantic information for multimodal binding. Concretely, GraphBind uses topology to organize self semantics and reliable neighborhood semantics into a global shared space that integrates structure and semantics, and adapts this space to discriminative and generative tasks through lightweight interfaces. Extensive experiments against 11 representative baselines demonstrate that GraphBind achieves leading performance on both discriminative and generative tasks, with relative improvements of up to 28.1% over the strongest baseline.

1 INTRODUCTION

Graph foundation models (GFMs) aim to learn generalizable graph representations through large-scale pretraining across graph domains and downstream tasks Liu et al. (2023a); Wang et al. (2025a), showing practical value in cross-domain graph classification and graph instruction tuning Liu et al. (2024); Tang et al. (2024). As real-world graphs increasingly contain heterogeneous modalities such as text and images Zhu et al. (2025); Wan et al. (2026), multimodal graph foundation models (MGFMs) have attracted growing attention in recent years. By integrating graph topology with multimodal attributes, MGFMs learn transferable representations for diverse graph tasks while supporting multimodal reasoning and generation He et al. (2025); Liu et al. (2026); Sun et al. (2026).

Despite substantial progress, learning reliable shared representations from graphs with incomplete node modalities remains challenging, which creates a critical data bottleneck for large-scale MAG pretraining. This bottleneck can be summarized from three perspectives: (1) **Incomplete Pretraining Corpora.** Real-world MAGs often contain incomplete node attributes. Restricting pretraining to nodes with complete observations substantially limits the scale and diversity of usable graph corpora Wu et al. (2026); Wang et al. (2023). (2) **Fragmented Semantic Spaces.** Separate modality encoding and alignment may produce fragmented semantic spaces, making it difficult to construct the unified representation space required by MGFMs. (3) **Limited Topological Binding.** Existing MGFMs primarily incorporate graph topology as structural context, with limited exploration of its role in guiding multimodal binding and shaping a unified representation space He et al. (2025); Liu et al. (2026); Sun et al. (2026). Our empirical study in Sec. 3 shows that graph neighborhoods provide additional semantic signals beyond node-local modalities. This finding supports the use of topology as a stable reference when learning from pretraining corpora with incomplete modalities.

Building upon this observation, we propose **GraphBind**, a novel MGFM framework that binds heterogeneous modality semantics through graph topology into a unified representation space. The key insight of GraphBind is that graph topology provides a stable structural reference and useful neighborhood information for multimodal binding. Concretely, GraphBind consists of three coupled

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.

designs. (1) **Incomplete Corpus Pretraining.** GraphBind explicitly introduces modality absence into pretraining, enabling the model to fully utilize incomplete MAGs as pretraining corpora and improve its robustness. (2) **Topology-Driven Multimodal Binding.** GraphBind uses graph topology as structural guidance to organize node semantics, neighborhood context, and multimodal attributes into a unified representation space. (3) **Unified Adaptation.** GraphBind adapts the unified representation space to both discriminative and generative tasks through lightweight fine-tuning while preserving the shared semantic interface learned during pretraining.

Our Contributions: (1) **In-depth Investigation.** We revisit realistic modality absence in MAGs and empirically show that graph neighborhoods provide useful recovery signals beyond the available attributes of each node. (2) **Novel Framework.** We propose GraphBind, a topology-guided multimodal graph foundation model that integrates missing-modality pretraining, topology-driven semantic recovery, and unified representation adaptation. (3) **SOTA Performance.** Extensive experiments against eleven representative baselines demonstrate leading performance on discriminative and generative tasks, with relative improvements of up to 5.4% in MRR and 28.1% in BLEU-4, respectively.

2 PRELIMINARIES

2.1 PROBLEM FORMULATION

A multimodal attributed graph (MAG) is denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \{\mathbf{X}^{(m)}\}_{m\in\mathcal{M}}, \mathbf{A})$, where $\mathcal{V} = \{v_i\}_{i=1}^{N}$ is the node set, $\mathcal{E}$ is the edge set, $N = |\mathcal{V}|$, and $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the adjacency matrix. Each node v_i can have multiple modalities from the modality set $\mathcal{M}$, such as text and images. For modality $m \in \mathcal{M}$, $\mathbf{x}_i^{(m)} \in \mathbb{R}^{d_m}$ denotes the raw attribute of node v_i, and $\mathbf{X}^{(m)} \in \mathbb{R}^{N\times d_m}$ denotes the modality feature matrix.

Given a pretraining collection of MAGs $\mathbb{G}_{\text{pre}}$, an MGFM f_Θ learns generalizable graph representations through self-supervised pretraining across graphs. The pretrained model maps each node v_i to a unified representation $\mathbf{z}_i \in \mathbb{R}^d$ and is adapted with task-specific modules. We consider two task families. Discriminative tasks include node classification and link prediction. Generative tasks include Graph to Text Generation (G2Text) and Graph to Image Generation (G2Image). Detailed task descriptions are provided in Appendix A.6.

2.2 GRAPH FOUNDATION MODELS

Graph foundation models learn transferable representations through pretraining across graph domains and tasks Liu et al. (2023a); Wang et al. (2025a). Existing methods represent graph computation with task trees or transferable vocabularies Wang et al. (2025b; 2024), learn reusable graph tokens Xia et al. (2024), or model graphs in geometric spaces Sun et al. (2025). These methods support broad transfer across graph tasks, while their pretraining mainly focuses on graph structure and a single attribute space. MAGs additionally require a shared space that connects heterogeneous node attributes with topology.

2.3 MULTIMODAL GRAPH LEARNING

Multimodal graph learning combines graph topology with node attributes such as text and images Zhu et al. (2025); Wan et al. (2026). Earlier methods use graph convolution, attention, or global graph structures to fuse modalities for specific tasks Wei et al. (2019); Tao et al. (2020); Guo et al. (2024), while recent methods jointly update topology and modality representations or assign different roles to neighbors Zhu et al. (2026); Zuo et al. (2026). Multimodal graph foundation models further learn shared representations for transfer across datasets and tasks He et al. (2025); Liu et al. (2026). Together, these studies explore how node attributes and graph relations can be jointly modeled to support prediction within individual graphs and transfer across graph domains and tasks.

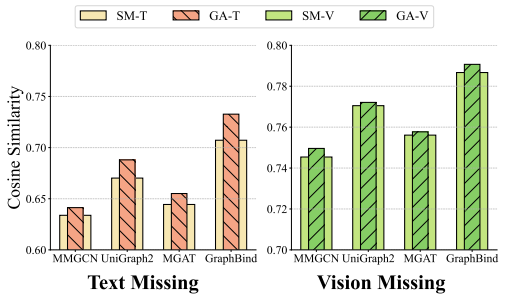


Figure 1: Comparison of recovery settings under single-modality absence on Grocery. *T* and *V* denotes text and vision recovery, respectively.

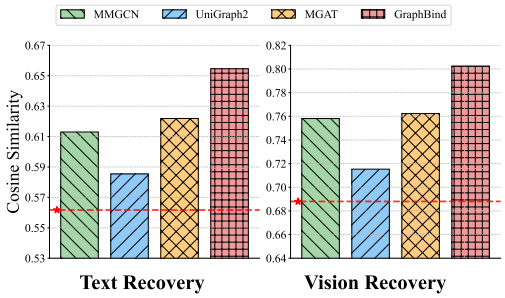


Figure 2: Graph-Assisted semantic recovery under both-modality absence on RedditS. Red lines denote the feature-mean prototype.

3 Empirical Investigation

In this section, we empirically examine why graph topology can serve as a structural reference for multimodal graph foundation modeling. We focus on two research questions: **Q1**: Are node-local modalities sufficient for multimodal binding? **Q2**: Can graph topology support modality recovery when neither modality of the center node is available?

Datasets, Methods, and Tasks. We conduct two recovery studies to analyze the semantic roles of node-local modalities and graph topology. We compare GraphBind with MMGCN, UniGraph2, and MGAT on modality feature recovery. Recovery quality is measured by the cosine similarity between the predicted feature and the corresponding reference feature. The first study is conducted on Grocery under single-modality absence, where either the textual or visual modality of the center node is unavailable. We compare two recovery settings with identical target-modality masks. *Single-Modality* (**SM**) recovery uses only the center node's remaining modality, with neighborhood messages and topology inputs disabled. *Graph-Assisted* (**GA**) recovery uses available features from one-hop neighbors. Neither setting receives the center node's target modality. The second study is conducted on RedditS when both modalities are absent, where both textual and visual modalities of the center node are unavailable. We use a feature-mean prototype as the non-structural reference and evaluate GA recovery for textual and visual semantics from surrounding nodes.

Individual node modalities provide limited semantic information (Answer for Q1). We first examine whether the remaining modality of a node provides sufficient information for semantic recovery. As shown in Fig. 1, SM consistently achieves lower recovery similarity than GA across all four methods under both absence settings. This observation indicates that the remaining modality provides useful but limited binding information. The limitation is especially clear when textual semantics need to be inferred from visual attributes, where the available visual information only partially reflects the missing textual semantics. Therefore, combining available node attributes with neighborhood information supports more effective multimodal binding in MAGs.

Graph topology provides shared semantic information (Answer for Q2). We further examine whether topology can provide semantic information when the center node has no available modality. As shown in Fig. 2, all graph-assisted methods achieve higher recovery similarity than the feature-mean prototype for both text and vision when neither modality of the center node is available. This result shows that graph neighborhoods preserve semantic information that is shared across nodes and transferable across target modalities. This observation suggests that graph topology can act as a structural reference for organizing surrounding node semantics in MAGs.

The two observations above motivate a topology-driven approach to multimodal graph foundation modeling. The available modalities provide useful but limited information for recovering missing modalities, as different modalities capture different aspects of node semantics. In comparison, graph topology provides a structural context for integrating multimodal information across connected nodes. These findings suggest that graph topology can serve as a structural reference for integrating multimodal information across nodes into shared semantic representations. This supports MGFM pretraining on graph corpora with incomplete node modalities.

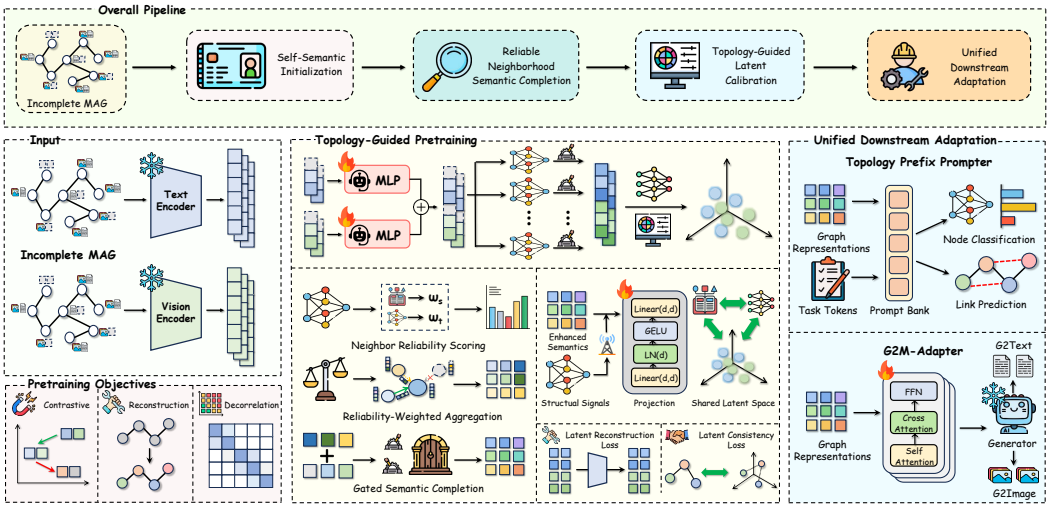


Figure 3: The overall framework of GraphBind.

4 METHODOLOGY

In this section, we present **GraphBind**, a multimodal graph foundation model that uses graph topology as a structural reference to bind node modalities into a unified representation space. As illustrated in Fig. 3, we first simulate realistic modality absence during pretraining, then construct a topology-guided shared representation through self-semantic initialization, reliable neighborhood aggregation, and topology-guided latent calibration, and finally adapt the learned shared space to discriminative and generative tasks with lightweight task interfaces.

4.1 MODALITY ABSENCE SIMULATION

Large-scale MAG pretraining cannot rely only on nodes with complete modalities, since naturally collected graph corpora have uneven modality availability. GraphBind therefore controls modality availability for each node, allowing incomplete MAGs to participate in pretraining.

Inspired by Wu et al. (2026), we simulate realistic modality absence during pretraining. Specifically, we consider three simulation protocols: *Independent Absence* randomly removes each modality. *Neighborhood Correlated Absence* removes modalities from adjacent nodes to represent local acquisition failure. *Property Dependent Absence* sets the absence probability according to node properties such as degree or community membership. The input availability received by the encoder is defined as

$$\tilde{a}_i^{(m)} = a_i^{(m)}(1 - s_i^{(m)})(1 - r_i^{(m)}), \tag{1}$$

where $a_i^{(m)}$, $s_i^{(m)}$, and $r_i^{(m)}$ indicate original availability, simulated absence, and reconstruction masking. The encoder receives only observed and unmasked modalities, while the original features remain supervision targets. Sampling details are provided in the Appendix A.2.

4.2 TOPOLOGY-GUIDED PRETRAINING

Incomplete Self-Semantic Initialization. Under modality absence, the remaining attributes of a node form an incomplete semantic state. We extract modality features with frozen encoders, project them into a common hidden space, and average only the modalities visible to the encoder:

$$\hat{\mathbf{x}}_i^{(m)} = P_m(\mathbf{x}_i^{(m)}), \quad \mathbf{h}_i^{(0)} = \frac{\sum_{m\in\mathcal{M}} \tilde{a}_i^{(m)} \hat{\mathbf{x}}_i^{(m)}}{\max(1, \sum_{m\in\mathcal{M}} \tilde{a}_i^{(m)})}, \tag{2}$$

where P_m is the modality projector and $\mathbf{h}_i^{(0)}$ is the initial self semantic state. For intuition, Eq. 2 initializes node v_i from the semantic information available under the current absence pattern. The averaging operation removes scale differences caused by varying numbers of observed modalities.

Reliable Neighborhood Semantic Completion. Node-local observations alone may provide insufficient information for multimodal binding. Although graph neighborhoods can supply missing semantics, not every neighbor provides relevant information. We therefore use topology to score and select reliable neighbors before aggregation. Let $\mathbf{t}_i = f_{\text{topo}}(\mathbf{p}_i)$ be the topology encoding of node v_i. For an edge $(j \to i)$, GraphBind scores neighbor reliability from self semantics, neighbor semantics, topology encodings, their interactions, and local edge strength:

$$r_{ij} = \sigma(\mathbf{w}_r^\top \boldsymbol{\phi}_{ij}), \quad \boldsymbol{\phi}_{ij} = \Phi(\mathbf{h}_i^{(0)}, \mathbf{h}_j^{(0)}, \mathbf{t}_i, \mathbf{t}_j, \omega_{ij}), \tag{3}$$

where ω_{ij} is a normalized diffusion weight, $\Phi(\cdot)$ concatenates and projects the input features through a GELU layer, and r_{ij} measures whether neighbor v_j is useful for node v_i. The reliable message and completed state are computed as

$$\begin{aligned} \alpha_{ij} &= \frac{r_{ij}}{\sum_{k \in \mathcal{N}(i)} r_{ik} + \epsilon}, \quad \mathbf{m}_i = \sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W}_n \mathbf{h}_j^{(0)}, \\ \mathbf{s}_i &= \mathbf{g}_i \odot \mathbf{W}_s \mathbf{h}_i^{(0)} + (1 - \mathbf{g}_i) \odot \mathbf{m}_i, \end{aligned} \tag{4}$$

where $\mathbf{g}_i$ is a learned feature-wise gate balancing self and neighborhood semantics, and $\mathbf{s}_i$ is the completed semantic state. The reliability weights prioritize neighbors that provide relevant semantic and structural information.

Topology-Guided Shared Latent Calibration. The completed state combines local node and neighborhood semantics, but it lacks global structural information. Therefore, we project it into a shared latent space and use topology to generate feature-wise calibration parameters:

$$\mathbf{u}_i = P_{\text{sh}}(\mathbf{s}_i), \quad [\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i] = C_{\text{topo}}(\mathbf{t}_i), \quad \mathbf{z}_i^{\text{anc}} = \boldsymbol{\gamma}_i \odot \mathbf{u}_i + \boldsymbol{\beta}_i, \tag{5}$$

where P_{sh} projects the completed state into the shared space and C_{topo} produces calibration parameters. This calibration places node semantics in a common coordinate system. To prevent this projection from discarding useful local semantics, we add two lightweight constraints over reliable neighbors $R(i)$ selected by their reliability scores:

$$\begin{aligned} \mathcal{L}_{\text{lrec}} &= \frac{1}{N} \sum_i \| D_{\text{sh}}(\mathbf{u}_i) - \text{sg}(\mathbf{s}_i) \|_2^2, \\ \mathcal{L}_{\text{lcons}} &= \sum_i \sum_{j \in R(i)} r_{ij} \| \Delta \mathbf{z}_{ij}^{\text{anc}} - \Delta \mathbf{s}_{ij} \|_2^2, \end{aligned} \tag{6}$$

where $\text{sg}(\cdot)$ denotes stop-gradient, $\Delta \mathbf{z}_{ij}^{\text{anc}} = \mathbf{z}_i^{\text{anc}} - \mathbf{z}_j^{\text{anc}}$, and $\Delta \mathbf{s}_{ij} = \mathbf{s}_i - \mathbf{s}_j$. The reconstruction constraint retains completed semantics after projection, and the consistency constraint keeps local geometry stable after calibration. Together, they ensure that topology organizes multimodal semantics without collapsing informative local variation. Due to space limitations, the complete pretraining objective and individual loss terms are detailed in Appendix A.2.

4.3 Unified Downstream Adaptation

Based on the shared space constructed during pretraining, we aim to adapt GraphBind to different downstream tasks through lightweight tuning without disrupting the pretrained representation geometry. For discriminative tasks, we design a topology prefix prompter to perform task-specific geometric adjustment. For generative tasks, the G2M-Adapter converts shared graph representations into continuous token embeddings for language and image generators.

Topology Prefix Prompter. For discriminative tasks, different objectives such as node classification and link prediction may require different local decision geometries. Inspired by continuous prompt learning Zhou et al. (2022) and graph prompt tuning Liu et al. (2023b); Sun et al. (2023), we introduce a shared prompt bank $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_K]^\top$ and a task token $\mathbf{c}_s$. The prompt bank provides

reusable task context, while $\mathbf{c}_s$ selects the context needed by a specific task:

$$b_{ik}^{(s)} = \mathrm{Attn}([\mathbf{z}_i^{\mathrm{anc}}; \mathbf{c}_s], \mathbf{p}_k), \quad \bar{\mathbf{p}}_i^{(s)} = \sum_{k=1}^{K} \mathrm{Softmax}_k(b_{ik}^{(s)})\mathbf{p}_k, \tag{7}$$

The prompted representation is obtained by adding the selected prompt context and task token to $\mathbf{z}_i^{\mathrm{anc}}$, followed by a small residual update. This adapter changes the representation only through a lightweight task-specific path, allowing the downstream head to specialize while the pretrained shared space remains stable.

G2M-Adapter. Generative tasks require graph side information to be expressed as context tokens compatible with pretrained language or visual generators. This interface should retain the node semantics and neighborhood information learned during pretraining. Inspired by lightweight modality bridging Li et al. (2023), we collect the target node and its one-hop neighbors into a graph context sequence $\mathbf{Z}_{\mathrm{topo}}$. Gated multihead attention converts their shared representations into tokens for the generator:

$$\mathbf{E}_s = F_s(\mathbf{Z}_{\mathrm{topo}}), \quad \bar{\mathbf{Z}}_s = O_s(\mathbf{Q}_s + \alpha_s \,\mathrm{MHA}(\mathbf{Q}_s, \mathbf{E}_s, \mathbf{E}_s)), \tag{8}$$

where s denotes text or image generation, and F_s and O_s denote the graph context encoder and output token mapping, respectively. $\mathbf{Q}_s$ contains learnable queries for text and equals $\mathbf{E}_s$ for images; α_s is a scalar gate initialized to zero. For intuition, attention allows each query to aggregate relevant node information, while the gate controls the contribution of the aggregated context. The resulting tokens serve as language input prefixes or diffusion model inputs, allowing the generator to use both target-node semantics and neighborhood information. This design reduces the representation gap between continuous graph features and the token interface required by the generator.

4.4 Theoretical Analysis

4.4.1 How Do Reliable Neighbors Improve Robustness?

Theorem 4.1 (Reliable Neighbor Robustness)**.** *Let $R(i)$ and $U(i) = \mathcal{N}(i) \setminus R(i)$ denote the reliable and less reliable neighbors of node i. Suppose their scores satisfy $r_{ij} \geq r_+$ on $R(i)$ and $r_{ij} \leq r_-$ on $U(i)$, where $r_+ > r_-$. Consider two completion inputs with the same center-node state and reliability scores that differ only in the transformed representations of neighbors in $U(i)$, each by at most δ. Their completed states satisfy*

$$\|\mathbf{s}_i - \mathbf{s}_i'\|_2 \leq \frac{|U(i)|r_-}{|R(i)|r_+ + |U(i)|r_- + \epsilon}\,\delta. \tag{9}$$

Theorem 4.1 formalizes the filtering role of reliability scoring. Reliability-weighted aggregation prevents less reliable neighbors from dominating the completion process, allowing topology to supply complementary semantics without making the completed state overly sensitive to irrelevant or noisy neighborhood information. This result explains why GraphBind estimates neighbor reliability in Eq. 3 before performing semantic completion in Eq. 4.

4.4.2 How Is Shared Geometry Preserved?

Theorem 4.2 (Shared Geometry Preservation)**.** *Let $W_R = \sum_i \sum_{j \in R(i)} r_{ij} > 0$. When i is sampled uniformly from $\mathcal{V}$, the latent constraints ensure for any $\kappa \in (0,1)$ that*

$$\begin{gathered} \Pr_i\Big[\|D_{\mathrm{sh}}(\mathbf{u}_i) - \mathbf{s}_i\|_2 > \sqrt{\mathcal{L}_{\mathrm{lrec}}/\kappa}\Big] \leq \kappa, \\ \left(\frac{1}{W_R}\sum_i \sum_{j \in R(i)} r_{ij}\left|\|\Delta\mathbf{z}_{ij}^{\mathrm{anc}}\|_2 - \|\Delta\mathbf{s}_{ij}\|_2\right|^2\right)^{1/2} \leq \sqrt{\mathcal{L}_{\mathrm{lcons}}/W_R}. \end{gathered} \tag{10}$$

The first result converts $\mathcal{L}_{\mathrm{lrec}}$ into a node-level guarantee. A small reconstruction loss limits the fraction of nodes whose projected states cannot recover their completed semantics. The second result bounds the weighted average change in reliable edge lengths after calibration through $\mathcal{L}_{\mathrm{lcons}}$. Complete proofs of both theorems are provided in Appendix A.1.

Table 1: Model performance in discriminative tasks. The best results are **bold**. The second-best results are *underlined*.

Methods	Node Classification				Link Prediction				Few-Shot Node Classification			
	Toys		Grocery		DY		Bili_Dance		Movies-5Way		RedditS-5Way	
	Acc	macro-F1	Acc	macro-F1	MRR	Hits@3	MRR	Hits@3	5-Shot	3-Shot	5-Shot	3-Shot
MMGCN	$76.54_{\pm 0.25}$	$73.61_{\pm 0.36}$	$79.27_{\pm 0.20}$	$68.78_{\pm 0.43}$	$66.67_{\pm 0.24}$	$69.62_{\pm 0.26}$	$65.47_{\pm 0.57}$	$70.72_{\pm 0.50}$	$46.79_{\pm 4.22}$	$42.42_{\pm 4.15}$	$\underline{95.64}_{\pm 0.87}$	$94.83_{\pm 0.71}$
MGAT	$76.82_{\pm 0.27}$	$68.23_{\pm 0.52}$	$80.29_{\pm 0.22}$	$65.14_{\pm 0.54}$	$65.45_{\pm 0.25}$	$68.51_{\pm 0.25}$	$65.48_{\pm 0.82}$	$69.95_{\pm 0.36}$	$47.45_{\pm 3.47}$	$42.39_{\pm 3.42}$	$94.48_{\pm 1.06}$	$93.89_{\pm 0.57}$
LGMRec	$77.72_{\pm 0.21}$	$75.60_{\pm 0.08}$	$81.29_{\pm 0.03}$	$70.85_{\pm 0.60}$	$67.89_{\pm 0.12}$	$72.61_{\pm 0.17}$	$69.35_{\pm 0.26}$	$75.73_{\pm 0.52}$	$\underline{53.85}_{\pm 4.19}$	$49.49_{\pm 4.30}$	$95.60_{\pm 0.86}$	$\underline{94.98}_{\pm 0.62}$
TMTE	$77.74_{\pm 0.25}$	$70.60_{\pm 0.46}$	$78.61_{\pm 0.27}$	$60.40_{\pm 1.08}$	$19.71_{\pm 1.80}$	$19.97_{\pm 1.87}$	$45.32_{\pm 1.41}$	$58.83_{\pm 1.41}$	$50.60_{\pm 4.15}$	$46.78_{\pm 4.70}$	$95.30_{\pm 0.90}$	$94.40_{\pm 0.73}$
RoleMAG	$\mathbf{78.85}_{\pm 0.32}$	$\mathbf{77.23}_{\pm 0.38}$	$\underline{81.71}_{\pm 0.34}$	$\underline{74.09}_{\pm 0.38}$	$69.16_{\pm 1.04}$	$75.92_{\pm 1.05}$	$64.11_{\pm 1.25}$	$75.85_{\pm 1.36}$	$52.92_{\pm 2.54}$	$46.72_{\pm 3.97}$	$95.41_{\pm 0.80}$	$94.64_{\pm 0.69}$
GIT	$59.88_{\pm 0.43}$	$55.31_{\pm 1.98}$	$65.69_{\pm 0.86}$	$52.07_{\pm 0.81}$	$61.34_{\pm 0.88}$	$66.39_{\pm 0.70}$	$71.63_{\pm 1.00}$	$74.12_{\pm 0.49}$	$27.98_{\pm 2.70}$	$26.58_{\pm 2.56}$	$50.40_{\pm 4.13}$	$47.69_{\pm 3.26}$
GFT	$76.45_{\pm 0.27}$	$73.78_{\pm 0.35}$	$80.18_{\pm 0.48}$	$70.65_{\pm 0.76}$	$\underline{72.79}_{\pm 0.35}$	$\mathbf{76.30}_{\pm 0.07}$	$\underline{72.01}_{\pm 0.41}$	$\underline{77.86}_{\pm 0.31}$	$44.36_{\pm 3.06}$	$40.69_{\pm 3.09}$	$90.58_{\pm 1.60}$	$89.15_{\pm 1.66}$
OpenGraph	$54.80_{\pm 0.56}$	$45.52_{\pm 1.01}$	$67.97_{\pm 1.03}$	$49.03_{\pm 2.27}$	$58.22_{\pm 0.45}$	$65.31_{\pm 0.44}$	$54.72_{\pm 0.35}$	$64.80_{\pm 0.41}$	$41.67_{\pm 4.59}$	$38.80_{\pm 3.59}$	$76.92_{\pm 5.11}$	$75.25_{\pm 5.01}$
RiemannGFM	$77.39_{\pm 0.19}$	$75.74_{\pm 0.19}$	$81.16_{\pm 0.32}$	$71.93_{\pm 0.16}$	$27.30_{\pm 0.75}$	$31.31_{\pm 1.07}$	$39.45_{\pm 0.44}$	$48.20_{\pm 0.82}$	$43.97_{\pm 4.01}$	$39.84_{\pm 4.61}$	$94.83_{\pm 0.72}$	$94.22_{\pm 0.67}$
UniGraph2	$73.66_{\pm 0.89}$	$67.25_{\pm 2.24}$	$70.32_{\pm 2.39}$	$51.36_{\pm 4.65}$	$63.87_{\pm 0.83}$	$70.64_{\pm 0.98}$	$50.60_{\pm 7.37}$	$59.01_{\pm 9.25}$	$52.40_{\pm 2.90}$	$\underline{49.67}_{\pm 3.39}$	$84.53_{\pm 5.29}$	$83.67_{\pm 5.52}$
PLANET	$77.57_{\pm 0.38}$	$75.13_{\pm 0.32}$	$75.54_{\pm 2.26}$	$59.94_{\pm 5.49}$	$68.75_{\pm 0.37}$	$71.78_{\pm 0.27}$	$70.28_{\pm 0.95}$	$73.86_{\pm 0.62}$	$42.23_{\pm 5.93}$	$44.63_{\pm 2.03}$	$65.77_{\pm 6.97}$	$64.33_{\pm 6.34}$
GraphBind	$\underline{78.08}_{\pm 0.05}$	$\underline{77.06}_{\pm 0.04}$	$\mathbf{81.86}_{\pm 0.12}$	$\mathbf{74.58}_{\pm 0.14}$	$\mathbf{75.23}_{\pm 0.70}$	$\underline{76.00}_{\pm 0.41}$	$\mathbf{75.88}_{\pm 0.37}$	$\mathbf{79.36}_{\pm 0.32}$	$\mathbf{56.44}_{\pm 1.69}$	$\mathbf{52.67}_{\pm 1.88}$	$\mathbf{96.37}_{\pm 0.75}$	$\mathbf{95.34}_{\pm 0.67}$

5 EXPERIMENTS

In this section, we first describe the experimental setup. Then, we conduct comprehensive evaluations to answer the following research questions: **Q1**: How does GraphBind perform against representative methods on discriminative and generative tasks? **Q2**: What are the contributions of the main components in GraphBind? **Q3**: How robust is GraphBind under varying modality absence patterns? **Q4**: What are the computational and storage costs of GraphBind?

5.1 EXPERIMENTAL SETUP

Datasets. We evaluate GraphBind on eight representative MAG datasets from the OpenMAG benchmark Wan et al. (2026) covering e-commerce, social media, video, image-text, and art domains. Specifically, Grocery, Movies, and Toys Ni et al. (2019), together with RedditS Desai et al. (2021), are used for node classification; DY and Bili_Dance Zhang et al. (2024) for link prediction; Flickr30k Plummer et al. (2015) for graph-to-text generation; and SemArt Garcia & Vogiatzis (2018) for graph-to-image generation. Detailed dataset statistics are provided in the Appendix A.4.

Baselines. We consider three categories of representative methods: (1) *MAG models*: MMGCN Wei et al. (2019), MGAT Tao et al. (2020), LGMRec Guo et al. (2024), TMTE Zhu et al. (2026), and RoleMAG Zuo et al. (2026); (2) *GFMs*: GIT Wang et al. (2025b), GFT Wang et al. (2024), OpenGraph Xia et al. (2024), and RiemannGFM Sun et al. (2025); and (3) *MGFMs*: UniGraph2 He et al. (2025) and PLANET Liu et al. (2026). Detailed descriptions and implementation settings for all baselines are provided in the Appendix A.5.

Downstream Tasks. We evaluate GraphBind on two task families: (1) *Discriminative tasks*, including node classification and link prediction; and (2) *Generative tasks*, including graph-to-text generation (G2Text) and graph-to-image generation (G2Image). Detailed task protocols and evaluation metrics are provided in the Appendix A.6.

5.2 MAIN RESULTS (ANSWER FOR Q1)

To answer Q1, we compare GraphBind with representative MAG models, GFMs, and MGFMs on discriminative and generative tasks in Tables 1 and 2.

Discriminative Tasks. Under supervised tasks, GraphBind achieves the best results on Grocery and Bili_Dance and the highest MRR on DY, while ranking second on Toys and in Hits@3 on DY. It also achieves the best performance across all four few-shot settings, improving accuracy by 2.59 percentage points over the strongest baseline on Movies-5Way 5-shot. These results demonstrate that binding structural and semantic information into a global shared space yields task-relevant representations that transfer effectively across supervision regimes. Furthermore, the topology prefix prompter provides lightweight task-specific adaptation to this space, enabling GraphBind to specialize for different discriminative objectives.

Table 2: Model performance in generative tasks. The best results are **bold**. The second results are *underlined*.

Methods	G2Image		G2Text	
	CLIP-S	DINO-S	BLEU-4	ROUGE-L
MMGCN	$69.26_{\pm 0.06}$	$46.60_{\pm 0.03}$	$7.36_{\pm 0.12}$	$30.85_{\pm 0.10}$
MGAT	$67.65_{\pm 0.01}$	$45.23_{\pm 0.11}$	$7.30_{\pm 0.05}$	$30.86_{\pm 0.15}$
LGMRec	$68.69_{\pm 0.07}$	$47.25_{\pm 0.05}$	$7.54_{\pm 0.07}$	$31.42_{\pm 0.08}$
UniGraph2	$\underline{69.78}_{\pm 0.06}$	$\mathbf{49.22}_{\pm 0.08}$	$5.57_{\pm 0.25}$	$26.87_{\pm 0.65}$
PLANET	$66.76_{\pm 0.03}$	$45.32_{\pm 0.16}$	$\underline{12.83}_{\pm 0.73}$	$\underline{38.95}_{\pm 1.10}$
GraphBind	$\mathbf{70.01}_{\pm 0.02}$	$\underline{48.91}_{\pm 0.14}$	$\mathbf{16.43}_{\pm 0.36}$	$\mathbf{43.45}_{\pm 0.10}$

Table 3: Ablation results on Toys, DY, SemArt, and Flickr30k for the 4 tasks, respectively. The best results are **bold**.

Variant	Downstream Tasks			
	NC	LP	G2Image	G2Text
GCN	$75.27_{\pm 0.12}$	$74.91_{\pm 1.10}$	$64.48_{\pm 0.62}$	$13.05_{\pm 0.05}$
GAT	$75.79_{\pm 0.73}$	$74.70_{\pm 1.03}$	$65.66_{\pm 2.37}$	$12.86_{\pm 0.07}$
SAGE	$75.83_{\pm 0.66}$	$75.19_{\pm 1.12}$	$62.75_{\pm 1.44}$	$12.91_{\pm 0.03}$
T.E.	$73.22_{\pm 1.73}$	$66.39_{\pm 1.29}$	$64.25_{\pm 1.79}$	$14.39_{\pm 0.08}$
N.O.	$73.33_{\pm 1.62}$	$69.73_{\pm 1.12}$	$66.65_{\pm 1.72}$	$14.67_{\pm 0.23}$
GraphBind	$\mathbf{77.06}_{\pm 0.04}$	$\mathbf{75.23}_{\pm 0.70}$	$\mathbf{70.01}_{\pm 0.02}$	$\mathbf{16.43}_{\pm 0.36}$

Generative Tasks. For G2Image, GraphBind achieves 70.01 CLIP-S and 48.91 DINO-S, delivering strong visual semantic alignment, while for G2Text, it obtains 16.43 BLEU-4 and 43.45 ROUGE-L, surpassing the strongest baseline by 3.60 and 4.50 points, respectively. These results indicate that the shared representation provides complementary node semantics and neighborhood information for generation. Furthermore, the G2Text gains indicate that the G2M-Adapter enables generators to effectively use this graph context through lightweight adaptation.

5.3 Ablation Study (Answer for Q2)

To address Q2, we compare GraphBind with 5 variants to examine its core architecture, which couples reliable neighborhood semantic completion with topology guided shared latent calibration. Since these stages are coupled, each variant replaces them together under the same experimental protocol. We organize the variants into three groups: (1) *Simple Propagation*: We apply standard GNN encoders *GCN* Kipf & Welling (2017), *GAT* Veličković et al. (2018), and *GraphSAGE* Hamilton et al. (2017) over the original adjacency, retaining neighborhood propagation without neighbor reliability estimation, the self retention gate, or topology conditioned calibration; (2) *Topology Encoding (T.E.)*: We apply a node-wise semantic transformation using an MLP with both node semantics and topology encodings, without neighborhood propagation; and (3) *Node Only (N.O.)*: We remove topology encodings and the entire graph aware path, directly using the initialized node semantics.

As shown in Table 3, GraphBind achieves the best result on all four tasks, with larger gains over the strongest variants on the two generative tasks. Among the variants, *Simple Propagation* performs better on Toys and DY, while *N.O.* performs best on SemArt and Flickr30k. This task-dependent ordering indicates that direct propagation over the original adjacency provides useful information for discriminative graph tasks and can introduce unsuitable context for generation. Moreover, *T.E.* performs below *N.O.* on every dataset, showing that adding a node-wise transformation and topology encoding without neighborhood propagation does not improve upon self semantics alone. Overall, the proposed topology-guided encoder outperforms the tested alternatives across all four tasks, supporting its effectiveness in learning shared graph representations.

5.4 Robustness Analysis (Answer for Q3)

To answer Q3, we evaluate robustness to modality absence on Grocery and RedditS by masking modality inputs at inference. Specifically, we impose two noise patterns to model distinct failure processes. *Independent* noise masks each node's modalities separately and simulates dispersed random failures. *Neighborhood* noise masks locally concentrated nodes and simulates correlated failures within a community or collection context. We increase the missing ratio from 0.0 to 0.8 in steps of 0.1 and report macro-F1. As shown in Fig. 4, performance declines for all methods as modality absence increases. Additionally, GraphBind exhibits the smoothest degradation and remains the leading method across both datasets and missing patterns. Under neighborhood noise, GraphBind retains leading performance even when correlated modality failures remove locally shared information, showing that it remains effective when modality absence is concentrated in neighboring nodes. Together, the results support the robustness of GraphBind to increasing modality absence under noisy settings.

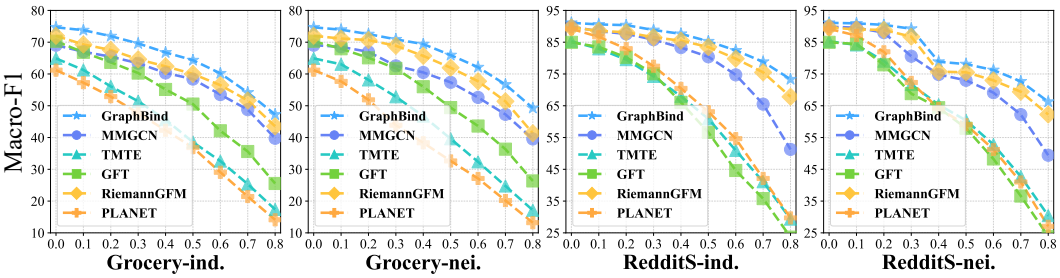


Figure 4: Test macro-F1 under independent and neighborhood modality absence on Grocery and RedditS. Missing ratios range from 0.0 to 0.8 in steps of 0.1.

Table 4: Computational efficiency of pretraining and node classification adaptation.

Methods	PT (s/epoch)	FT (s/epoch)	IT (ms/1k nodes)	Train. Para. (M)	Total Para. (M)
MMGCN	–	0.0319	12.19	3.3503	3.3503
MGAT	–	0.0389	30.63	2.5779	2.5779
LGMRec	–	0.0359	36.66	0.2405	0.2405
TMTE	–	0.1906	40.66	0.1180	0.1180
RoleMAG	–	0.4534	73.42	3.6890	3.6890
GIT	0.2755	0.0340	15.19	0.9905	0.9905
GFT	0.2344	0.0893	34.27	2.8273	2.8273
OpenGraph	0.1343	0.0950	39.64	1.1884	1.1884
RiemannGFM	1.3383	0.1891	33.26	0.2157	0.4717
UniGraph2	1.0177	0.1596	53.83	20.7237	20.7237
PLANET	1.6948	3.6859	92.39	221.9436	221.9436
GraphBind	0.3803	0.4022	59.17	5.3215	23.0692

5.5 Efficiency Analysis (Answer for Q4)

To answer Q4, we compare the computational costs of GraphBind with representative baselines in Table 4. GraphBind's finetuning and inference costs are higher than those of most lightweight MAG baselines, whereas its pretraining is faster than that of both MGFMs and its inference is faster than PLANET; this stage-dependent pattern indicates that GraphBind reduces repeated pretraining cost while incurring additional downstream cost for processing richer graph context. Moreover, among MGFMs, GraphBind requires substantially fewer trainable parameters, updating 5.32M parameters compared with 20.72M for UniGraph2 and 221.94M for PLANET, respectively. It also achieves the lowest proportion of trainable parameters among all compared methods. Combined with its strong classification performance, these findings demonstrate the effectiveness and parameter efficiency of GraphBind's lightweight finetuning design.

6 Conclusion

In this paper, we propose GraphBind, a multimodal graph foundation model that uses graph topology to bind heterogeneous modality semantics into a unified representation space. We revisit the role of topology in multimodal binding and show that existing methods largely treat it as a propagation or alignment signal, underexploiting its capacity to organize shared semantics. GraphBind couples topology-guided pretraining with lightweight fine-tuning, learning a shared space that integrates structure and semantics for both discriminative and generative tasks. Extensive experiments against 11 representative baselines show that GraphBind achieves strong performance across both discriminative and generative tasks, demonstrating strong generalization and robustness. We believe this work provides a valuable foundation for topology-driven multimodal graph foundation models and motivates further exploration of scalable graph-semantic binding.

AI USAGE STATEMENT

No generative AI tools were used for tasks requiring disclosure under the ICLR 2027 policy. The authors take full responsibility for the content of this paper.

REFERENCES

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Vijay Prakash Dwivedi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Graph neural networks with learnable structural and positional representations. In *International Conference on Learning Representations*, 2022.

Vijay Prakash Dwivedi, Chaitanya K. Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks. *Journal of Machine Learning Research*, 24(43):1–48, 2023.

Noa Garcia and George Vogiatzis. How to read paintings: Semantic art understanding with multimodal retrieval. In *European Conference on Computer Vision Workshops*, pp. 676–691, 2018.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Zhiqiang Guo, Jianjun Li, Guohui Li, Chaoyang Wang, Si Shi, and Bin Ruan. Lgmrec: Local and global graph learning for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8454–8462, 2024.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NIPS*, pp. 1024–1034, 2017.

Yufei He, Yuan Sui, Xiaoxin He, Yue Liu, Yifei Sun, and Bryan Hooi. UniGraph2: Learning a unified embedding space to bind multimodal graphs. In *Proceedings of the ACM Web Conference*, pp. 1759–1770, 2025. doi: 10.1145/3696410.3714818.

Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation: Benchmarking LLMs as semantic encoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3251–3265, 2026. doi: 10.18653/v1/2026.acl-long.147.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR (Poster)*. OpenReview.net, 2017.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. One for all: Towards training one graph model for all classification tasks. In *International Conference on Learning Representations*, 2024.

Jiawei Liu, Cheng Yang, Zhiyuan Lu, Junze Chen, Yibo Li, Mengmei Zhang, Ting Bai, Yuan Fang, Lichao Sun, Philip S. Yu, and Chuan Shi. Towards graph foundation models: A survey and beyond. *arXiv preprint arXiv:2310.11829*, 2023a.

Sicheng Liu, Xunkai Li, Daohan Su, Ru Zhang, Hongchao Qin, Ronghua Li, and Guoren Wang. Toward effective multimodal graph foundation model: A divide-and-conquer based approach. In *Proceedings of the International Conference on Machine Learning*, 2026.

Zemin Liu, Xingtong Yu, Yuan Fang, and Xinming Zhang. Graphprompt: Unifying pre-training and downstream tasks for graph neural networks. In *Proceedings of the ACM Web Conference*, 2023b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Jianmo Ni, Jiacheng Li, and Julian J. McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pp. 188–197, 2019.

Maxime Oquab, Timothée Darcet, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002. doi: 10.3115/1073083.1073135.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2641–2649, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Li Sun, Zhenhao Huang, Suyang Zhou, Qiqi Wan, Hao Peng, and Philip S. Yu. Riemanngfm: Learning a graph foundation model from riemannian geometry. In *Proceedings of the ACM Web Conference*, pp. 1154–1165, 2025.

Xiangguo Sun, Hong Cheng, Jia Li, Bo Liu, and Jihong Guan. All in one: Multi-task prompting for graph neural networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.

Yuanfu Sun, Kang Li, Pengkang Guo, Jiajin Liu, and Qiaoyu Tan. Mario: Multimodal graph reasoning with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19219–19228, 2026.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. Mgat: Multimodal graph attention network for recommendation. *Information Processing & Management*, 57(5):102277, 2020.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, 2015.

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

Chenxi Wan, Xunkai Li, Yilong Zuo, Haokun Deng, Sihan Li, Bowen Fan, Hongchao Qin, Ronghua Li, and Guoren Wang. OpenMAG: A comprehensive benchmark for multimodal-attributed graph. In *Proceedings of the International Conference on Machine Learning*, 2026.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Zehong Wang, Zheyuan Zhang, Nitesh Chawla, Chuxu Zhang, and Yanfang Ye. Gft: Graph foundation model with transferable tree vocabulary. In *Advances in Neural Information Processing Systems*, 2024.

Zehong Wang, Zheyuan Liu, Tianyi Ma, Jiazheng Li, Zheyuan Zhang, Xingbo Fu, Yiyang Li, Zhengqing Yuan, Wei Song, Yijun Ma, Qingkai Zeng, Xiusi Chen, Jianan Zhao, Jundong Li, Meng Jiang, Pietro Liò, Nitesh Chawla, Chuxu Zhang, and Yanfang Ye. Graph foundation models: A comprehensive survey. *arXiv preprint arXiv:2505.15116*, 2025a.

Zehong Wang, Zheyuan Zhang, Tianyi Ma, Nitesh V. Chawla, Chuxu Zhang, and Yanfang Ye. Towards graph foundation models: Learning generalities across graphs via task-trees. In *Proceedings of the International Conference on Machine Learning*, 2025b.

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the ACM International Conference on Multimedia*, pp. 1437–1445, 2019.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *Transactions on Machine Learning Research*, 2026.

Lianghao Xia, Ben Kao, and Chao Huang. Opengraph: Towards open graph foundation models. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024.

Hengrui Zhang, Qitian Wu, Junchi Yan, David Wipf, and Philip S. Yu. From canonical correlation analysis to self-supervised graph neural networks. In *Advances in Neural Information Processing Systems*, 2021.

Jiaqi Zhang, Yu Cheng, Yongxin Ni, Yunzhu Pan, Zheng Yuan, Junchen Fu, Youhua Li, Jie Wang, and Fajie Yuan. Ninerec: A benchmark dataset suite for evaluating transferable recommendation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022.

Jing Zhu, Yuhang Zhou, Shengyi Qian, Zhongmou He, Tong Zhao, Neil Shah, and Danai Koutra. Mosaic of modalities: A comprehensive benchmark for multimodal graph learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14215–14224, 2025. doi: 10.1109/CVPR52734.2025.01326.

Yinlin Zhu, Xunkai Li, Di Wu, Wang Luo, Miao Hu, and Guocong Quan. TMTE: Effective multi-modal graph learning with task-aware modality and topology co-evolution. In *Proceedings of the ACM International Conference on Multimedia*, 2026.

Yilong Zuo, Xunkai Li, Zhihan Zhang, Ronghua Li, and Guoren Wang. RoleMAG: Learning neighbor roles in multimodal graphs. In *Proceedings of the ACM International Conference on Multimedia*, 2026.

A Appendix

A.1 Theorem Proofs

Proof of Theorem 4.1. Fix a node i and compare the two completion inputs specified in the theorem. We trace the perturbation through the neighborhood aggregation and the completion gate.

Fixed quantities and perturbation source. Write $\mathbf{y}_{ij} = \mathbf{W}_n \mathbf{h}_j^{(0)}$ and $\mathbf{y}'_{ij} = \mathbf{W}_n \mathbf{h}_j^{(0)\prime}$ for the transformed messages received from neighbor j. The two inputs share the center-node state, topology encoding, and reliability scores. The completion gate depends only on these quantities through the center-node state, topology encoding, and mean neighbor reliability, so it is identical for both inputs. Their normalization constant and aggregation weights are therefore also shared:

$$\begin{aligned} Z_i &= \sum_{k \in \mathcal{N}(i)} r_{ik} + \epsilon, \\ \alpha_{ij} &= \alpha'_{ij} = \frac{r_{ij}}{Z_i}. \end{aligned} \tag{11}$$

The reliable messages remain unchanged, so $\mathbf{y}_{ij} = \mathbf{y}'_{ij}$ for $j \in R(i)$. All nonzero message differences are confined to $U(i)$, where $\|\mathbf{y}_{ij} - \mathbf{y}'_{ij}\|_2 \le \delta$.

Perturbation of the neighborhood message. Let $\mathbf{m}_i$ and $\mathbf{m}'_i$ denote the two neighborhood messages in Eq. 4. Subtracting their weighted sums cancels every reliable-neighbor term. The remaining terms satisfy

$$\begin{aligned} \|\mathbf{m}_i - \mathbf{m}'_i\|_2 &= \left\| \sum_{j \in U(i)} \alpha_{ij} (\mathbf{y}_{ij} - \mathbf{y}'_{ij}) \right\|_2 \\ &\le \sum_{j \in U(i)} \alpha_{ij} \|\mathbf{y}_{ij} - \mathbf{y}'_{ij}\|_2 \\ &\le \frac{U_i^w}{R_i^w + U_i^w + \epsilon}\, \delta, \end{aligned} \tag{12}$$

where $R_i^w = \sum_{j \in R(i)} r_{ij}$ and $U_i^w = \sum_{j \in U(i)} r_{ij}$ are the total reliability masses assigned to the two neighbor groups.

Control of the less reliable mass. Define $f(R, U) = U/(R + U + \epsilon)$. Its two partial derivatives have fixed signs:

$$\frac{\partial f}{\partial R} = -\frac{U}{(R + U + \epsilon)^2} \le 0, \qquad \frac{\partial f}{\partial U} = \frac{R + \epsilon}{(R + U + \epsilon)^2} \ge 0. \tag{13}$$

The score separation in Theorem 4.1 gives $R_i^w \ge |R(i)| r_+$ and $U_i^w \le |U(i)| r_-$. Applying both inequalities according to the monotonicity in Eq. 13 yields

$$\frac{U_i^w}{R_i^w + U_i^w + \epsilon} \le \frac{|U(i)| r_-}{|R(i)| r_+ + |U(i)| r_- + \epsilon}. \tag{14}$$

This bound identifies the exact quantity controlled by reliability scoring: the normalized mass assigned to the perturbed neighbor group decreases as the separation between r_+ and r_- increases.

Propagation through the completion gate. The center-node contribution in Eq. 4 is identical for both inputs and therefore cancels. Since $\mathbf{g}_i \in [0, 1]^d$, element-wise multiplication by $1 - \mathbf{g}_i$ is nonexpansive under the Euclidean norm. Consequently,

$$\begin{aligned} \|\mathbf{s}_i - \mathbf{s}'_i\|_2 &= \|(1 - \mathbf{g}_i) \odot (\mathbf{m}_i - \mathbf{m}'_i)\|_2 \\ &\le \|\mathbf{m}_i - \mathbf{m}'_i\|_2 \\ &\le \frac{|U(i)| r_-}{|R(i)| r_+ + |U(i)| r_- + \epsilon}\, \delta. \end{aligned} \tag{15}$$

The final line is exactly Eq. 9. □

Proof of Theorem 4.2. The two inequalities in the theorem correspond to the two constraints in Eq. 6. We establish them separately and then state their joint implication for the calibrated shared space.

Semantic recoverability. For each node, define the nonnegative reconstruction error

$$E_i = \|D_{\mathrm{sh}}(\mathbf{u}_i) - \mathbf{s}_i\|_2^2. \tag{16}$$

Stop-gradient changes the backward derivative while leaving the forward value intact. The first definition in Eq. 6 therefore implies

$$\mathbb{E}_{i\sim\mathrm{Unif}(\mathcal{V})}[E_i] = \frac{1}{N}\sum_{i=1}^{N} E_i = \mathcal{L}_{\mathrm{lrec}}. \tag{17}$$

If $\mathcal{L}_{\mathrm{lrec}} = 0$, nonnegativity forces every E_i to be zero. Now consider $\mathcal{L}_{\mathrm{lrec}} > 0$. For any $\kappa \in (0, 1)$, Markov's inequality applied to Eq. 17 gives the complete probability chain

$$\begin{aligned} \Pr_i\left[\|D_{\mathrm{sh}}(\mathbf{u}_i) - \mathbf{s}_i\|_2 > \sqrt{\mathcal{L}_{\mathrm{lrec}}/\kappa}\right] &= \Pr_i[E_i > \mathcal{L}_{\mathrm{lrec}}/\kappa] \\ &\le \frac{\mathbb{E}_i[E_i]}{\mathcal{L}_{\mathrm{lrec}}/\kappa} \\ &= \kappa. \end{aligned} \tag{18}$$

This proves the node-level recoverability statement.

Reliable local geometry. For each directed reliable pair (i, j), define its displacement error and edge-length distortion as

$$\mathbf{e}_{ij} = \Delta\mathbf{z}_{ij}^{\mathrm{anc}} - \Delta\mathbf{s}_{ij}, \qquad d_{ij} = \left|\|\Delta\mathbf{z}_{ij}^{\mathrm{anc}}\|_2 - \|\Delta\mathbf{s}_{ij}\|_2\right|. \tag{19}$$

The reverse triangle inequality gives $d_{ij} \le \|\mathbf{e}_{ij}\|_2$. Since every reliability score is nonnegative, this pointwise relation can be squared, weighted, and summed without changing its direction:

$$\begin{aligned} \sum_i \sum_{j\in R(i)} r_{ij} d_{ij}^2 &\le \sum_i \sum_{j\in R(i)} r_{ij}\|\mathbf{e}_{ij}\|_2^2 \\ &= \sum_i \sum_{j\in R(i)} r_{ij}\|\Delta\mathbf{z}_{ij}^{\mathrm{anc}} - \Delta\mathbf{s}_{ij}\|_2^2 \\ &= \mathcal{L}_{\mathrm{lcons}}. \end{aligned} \tag{20}$$

The normalization constant $W_R = \sum_i \sum_{j\in R(i)} r_{ij}$ is strictly positive by assumption. Dividing Eq. 20 by W_R and taking square roots gives

$$\left(\frac{1}{W_R}\sum_i \sum_{j\in R(i)} r_{ij}\left|\|\Delta\mathbf{z}_{ij}^{\mathrm{anc}}\|_2 - \|\Delta\mathbf{s}_{ij}\|_2\right|^2\right)^{1/2} \le \sqrt{\frac{\mathcal{L}_{\mathrm{lcons}}}{W_R}}. \tag{21}$$

This is the second inequality in Eq. 10. Together with Eq. 18, it shows that the two latent constraints control complementary properties: semantic decoding error at the node level and metric distortion over reliable local relations. □

A.2 Method Details

Modality Absence Simulation. For Independent Absence, we sample the simulated absence indicator for each observed node and modality pair as

$$s_i^{(m)} \sim \mathrm{Bernoulli}\left(\rho_m(e)\right), \quad a_i^{(m)} = 1, \tag{22}$$

where $\rho_m(e)$ denotes the absence probability of modality m at epoch e. We then sample an additional reconstruction mask only from encoder-visible modalities,

$$r_i^{(m)} \sim \mathrm{Bernoulli}\left(\mu_m\right), \quad a_i^{(m)} = 1, \quad s_i^{(m)} = 0. \tag{23}$$

The encoder receives the effective availability in Eq. 1. The supervision set for modality reconstruction contains both sources of hidden observations,

$$\mathcal{T}^{(m)} = \left\{ i : a_i^{(m)} = 1, \quad s_i^{(m)} = 1 \text{ or } r_i^{(m)} = 1 \right\}. \tag{24}$$

Thus, simulated absence evaluates recovery from the remaining modalities and topology, while the reconstruction mask supplies additional self-supervised targets without exposing their features to the encoder.

For Neighborhood Correlated Absence, we first sample seed nodes $\mathcal{C}^{(m)}$ and mask the neighborhoods within K hops of the seeds.

$$s_i^{(m)} = \mathbb{I}\left[\min_{v_c \in \mathcal{C}^{(m)}} \mathrm{dist}_{\mathcal{G}}(v_i, v_c) \le K \right]. \tag{25}$$

The number of seeds is selected to match the target missing ratio. For Property Dependent Absence, a nonnegative node property g_i determines the absence probability.

$$\Pr\left(s_i^{(m)} = 1 \mid g_i \right) = \min\left(1, c_m(e)\psi_m(g_i)\right), \tag{26}$$

where $c_m(e)$ is chosen to attain the desired average missing ratio and ψ_m is a nonnegative scaling function. At inference, we remove the simulated indicators and use the observed availability $a_i^{(m)}$ directly.

Reliable Neighborhood Semantic Completion. The topology input $\mathbf{p}_i$ combines Laplacian positional features Dwivedi et al. (2023) and random walk Dwivedi et al. (2022) return probabilities computed on the symmetric binary graph. The former use the eigenvectors associated with the smallest normalized Laplacian eigenvalues, excluding the first eigenvector, while the latter use diagonal entries of successive powers of the transition matrix. The encoder f_{topo} maps these features through two linear layers with LayerNorm and GELU between them.

In Eq. 3, Φ concatenates the two semantic states, the two topology encodings, their respective element-wise products, and ω_{ij}, then applies a linear projection and GELU. The final affine scoring layer includes the bias b_r. For the row-normalized transition matrix $\mathbf{T}$, the diffusion matrix is $\mathbf{W} = \eta \sum_{k=0}^{L} (1-\eta)^k \mathbf{T}^k$, where η is the restart coefficient and L is the diffusion depth. The weight for $j \to i$ is $\omega_{ij} = W_{ji}/W_{jj}$, bounded below by a small positive constant. This normalizes the source-to-target connection by the source's self-diffusion. Protocols that omit diffusion use unit edge weights. In Eq. 4, all incoming neighbors contribute through normalized reliability scores. The retention gate is a sigmoid applied to an affine projection of $[\mathbf{h}_i^{(0)}; \mathbf{t}_i; \bar{r}_i]$, where $\bar{r}_i$ is their mean reliability, yielding a separate balance for each feature dimension.

Topology-Guided Shared Latent Calibration. In Eq. 5, P_{sh} combines a two-layer projection with a gated linear semantic residual. The projection uses LayerNorm and GELU between its linear layers, and a sigmoid gate computed from $\mathbf{s}_i$ controls the residual. The topology conditioner uses two linear layers separated by GELU to predict scale and shift offsets. The scale is parameterized as $\boldsymbol{\gamma}_i = \mathbf{1} + \Delta\boldsymbol{\gamma}_i$, and the output layer starts at zero, so the affine calibration initially preserves $\mathbf{u}_i$. The bypass gate $\mathbf{b}_i$ is a sigmoid applied to an affine projection of $[\mathbf{s}_i; \mathbf{t}_i]$, retaining completed semantics directly in the final representation.

For Eq. 6, D_{sh} is a two-layer decoder with GELU that reconstructs $\mathbf{s}_i$ from the shared state before topology calibration. Both constraints detach their completed-semantic targets. To construct $R(i)$, we rank incoming neighbors by reliability, retain a bounded number of the highest-scoring neighbors, and apply a reliability threshold. If none passes, the highest-scoring neighbor is retained with its continuous reliability weight. The implementation averages reconstruction errors over nodes and coordinates, and consistency errors over selected edges and coordinates. These are normalized versions of the losses in Eq. 6, whose unnormalized notation is used in the theoretical bounds.

Pretraining Objectives.

Contrastive Alignment. GraphBind uses contrastive learning Radford et al. (2021); Girdhar et al. (2023) to bind each shared representation to node-local modality semantics and structurally related

semantics. Multiple positives provide complementary semantic targets when a node's own modalities are incomplete. For a pretraining batch $\mathcal{B}$, let $\mathcal{V}_m = \{i \in \mathcal{B} : a_i^{(m)} = 1\}$ collect nodes whose original modality m is available. We compute the cosine similarity $q_{ij}^{(m)}$ between anchor i and the modality representation of node j from the original, unmasked input stream.

Besides the same node, the positive set $\mathcal{P}_{i,+}^{(m)}$ contains candidates with sufficiently high diffusion weight and cross-modal similarity, ranked by the product of their diffusion weight and nonnegative similarity. Here, $d_{ij} = W_{ij}/W_{ii}$ measures diffusion from anchor i to candidate j. Hard negatives $\mathcal{H}_i^{(m)}$ have high similarity and low diffusion weight, and are ranked by $q_{ij}^{(m)} - d_{ij}$. Each selection retains a bounded number of candidates. These hard negatives emphasize semantic distinctions that graph connections help resolve. Feature dropout additionally provides an augmented modality view, whose dot product with the normalized anchor is denoted by $\widetilde{q}_{ij}^{(m)}$. The objective is

$$\begin{aligned} A_i^{(m)} &= e^{q_{ii}^{(m)}/\tau} + e^{\widetilde{q}_{ii}^{(m)}/\tau} + \sum_{j \in \mathcal{P}_{i,+}^{(m)}} e^{q_{ij}^{(m)}/\tau}, \\ B_i^{(m)} &= \sum_{j \in \mathcal{V}_m} d_{ij} e^{q_{ij}^{(m)}/\tau} + \sum_{j \in \mathcal{V}_m} e^{\widetilde{q}_{ij}^{(m)}/\tau} + \lambda_{\text{hard}} \sum_{j \in \mathcal{H}_i^{(m)}} e^{q_{ij}^{(m)}/\tau}, \\ \mathcal{L}_{\text{ctr}}^{(m)} &= -\frac{1}{|\mathcal{V}_m|} \sum_{i \in \mathcal{V}_m} \log \frac{A_i^{(m)}}{B_i^{(m)}}. \end{aligned} \tag{27}$$

Here, $\tau > 0$ is the temperature and λ_{hard} controls hard-negative emphasis. The denominator is evaluated over all available candidates in chunks. Disabled augmentation or mining removes the corresponding terms, and unavailable diffusion weights are replaced by one. Summing over modalities gives $\mathcal{L}_{\text{ctr}} = \sum_m \mathcal{L}_{\text{ctr}}^{(m)}$. Aggregating positive scores allows their contributions to follow semantic agreement, while hard-negative emphasis encourages finer distinctions within the shared space.

Masked Modality Reconstruction. Contrastive alignment primarily preserves relations among samples and can omit modality details needed by generative tasks. Following masked graph autoencoding Hou et al. (2022), GraphBind reconstructs the hidden observations indexed by $\mathcal{T}^{(m)}$. We use the scaled cosine error

$$\delta_{\text{sce}}(\mathbf{u}, \mathbf{v}) = \left(1 - \frac{\mathbf{u}^\top \mathbf{v}}{\max(\|\mathbf{u}\|_2, \epsilon) \max(\|\mathbf{v}\|_2, \epsilon)}\right)^\gamma, \qquad \gamma \geq 1. \tag{28}$$

Cosine reconstruction matches the angular geometry of the modality features and removes feature magnitude from the reconstruction error. This focuses supervision on semantic direction, which MSE also couples to feature norms. The implementation uses $\gamma = 1$. The masked reconstruction loss is

$$\mathcal{L}_{\text{rec}}^{(m)} = \frac{1}{|\mathcal{T}^{(m)}|} \sum_{i \in \mathcal{T}^{(m)}} \delta_{\text{sce}} \left(D_m(\mathbf{z}_i^{\text{anc}}), \mathbf{x}_i^{(m)}\right), \tag{29}$$

where D_m is the modality decoder and $\mathcal{L}_{\text{rec}} = \sum_m \mathcal{L}_{\text{rec}}^{(m)}$. Reconstructing hidden attributes encourages the shared representation to retain modality information recovered through the semantic completion path.

Latent Decorrelation. Alignment and reconstruction can encode correlated information repeatedly across latent dimensions, reducing the effective capacity of the shared space. Let $\bar{\mathbf{Z}}$ and $\bar{\mathbf{V}}^{(m)}$ denote column-wise centered and standardized anchor and modality matrices. Their empirical correlation matrices are

$$\mathbf{C}^{(a)} = \frac{1}{|\mathcal{B}|} \bar{\mathbf{Z}}^\top \bar{\mathbf{Z}}, \qquad \mathbf{C}^{(m)} = \frac{1}{|\mathcal{V}_m|} (\bar{\mathbf{V}}^{(m)})^\top \bar{\mathbf{V}}^{(m)}. \tag{30}$$

Following graph decorrelation learning Zhang et al. (2021), we penalize their off-diagonal entries:

$$\mathcal{L}_{\text{decor}} = \frac{1}{d(d-1)} \left[\sum_{u \neq v} (\mathbf{C}_{uv}^{(a)})^2 + \sum_{m \in \mathcal{M}} \sum_{u \neq v} (\mathbf{C}_{uv}^{(m)})^2 \right], \tag{31}$$

where d is the representation dimension. Penalizing correlation across distinct coordinates encourages complementary features for multimodal binding.

Total Objective. We jointly optimize semantic alignment, modality recovery, and shared-space regularization through

$$\begin{aligned}\mathcal{L}_{\text{pre}} = \lambda_{\text{ctr}}\mathcal{L}_{\text{ctr}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{dec}}\mathcal{L}_{\text{decor}} \\ + \lambda_{\text{lr}}\mathcal{L}_{\text{lrec}} + \lambda_{\text{lc}}\mathcal{L}_{\text{lcons}},\end{aligned} \tag{32}$$

where the five coefficients balance the objectives. Together, these terms train the shared representation to support cross-modal alignment and missing-modality completion while preserving useful semantic variation.

Topology Prefix Prompter. In Eq. 7, Attn is a scaled dot product between a projected query and projected prompt keys. The query combines the shared node representation with a linearly projected task token, and the values are the original prompt vectors. The selected context, shared representation, and projected task token are added and normalized before a two-layer MLP with GELU produces the residual update. A feature-wise task gate, obtained by applying $\tanh$ to an affine projection of the projected task token, scales this update before it is added to $\mathbf{z}_i^{\text{anc}}$. The residual output layer is initialized to zero, giving an identity mapping at initialization. When no task token is supplied, the query uses a zero condition and the residual update is applied directly.

G2M-Adapter. For Eq. 8, the context sequence places the target node first, followed by its incoming one-hop neighbors from the task metadata. Repeated entries and the target itself are removed from the neighbor list, which is truncated in metadata order. Unused slots contain zero vectors and are masked as attention keys.

For text generation, F_s projects the shared representations and applies local message passing with bidirectional links from the target to its neighbors and between consecutive retained neighbors. Each block combines transformed self states and mean neighbor states through GELU, a residual connection, and LayerNorm. Slot embeddings distinguish context positions. Learnable queries aggregate these states through the gated attention in Eq. 8, and O_s applies LayerNorm. The resulting tokens are prepended to the language input embeddings, with prefix length determined by the query bank.

For image generation, F_s is a linear projection and $\mathbf{Q}_s = \mathbf{E}_s$ gives self-attention over context nodes. After the gated residual update, O_s adds slot embeddings, normalizes and projects the states, and zeros invalid slots. A final linear projection expands each slot into generator tokens followed by LayerNorm, retaining the padded slots in the output sequence. These tokens are concatenated with text embeddings and supplied to the diffusion U-Net through its existing cross-attention interface.

A.3 Pseudocode of GraphBind

We provide the overall procedure for GraphBind in Algorithm 1.

A.4 Dataset Details

In this section, we provide detailed statics of the datasets in Table 5. Texual descriptions are shown as follows.

Table 5: Statistics of the datasets used in GraphBind.

Dataset	Modalities	Nodes	Edges	Classes	Domain
Grocery	Text, Visual	17,074	171,340	20	E-Commerce
RedditS	Text, Visual	15,894	566,160	20	Social Media
Movies	Text, Visual	16,672	218,390	20	E-Commerce
Toys	Text, Visual	20,695	126,886	18	E-Commerce
Ele-fashion	Text, Visual	97,766	199,602	12	E-Commerce
Sports	Text, Visual	50,250	356,202	–	E-Commerce
Bili_Dance	Text, Visual	2,307	9,127	–	Video Recommendation
DY	Text, Visual	8,299	35,627	–	Video Recommendation
SemArt	Text, Visual	21,382	1,216,432	–	Art Network
Flickr30k	Text, Visual	31,783	181,151	–	Image Network

Algorithm 1: Overall Procedure of GraphBind

Input: Pretraining graphs $\mathbb{G}_{\text{pre}}$ with modality availability; maximum epochs E; downstream graph and training data for task s.
Output: Pretrained encoder f_Θ and adapted task model.

```
/* Initialization                                                    */
initialize model parameters and extract features with frozen modality encoders;
prepare topology features p_i and diffusion weights;
for e = 1 to E do
    foreach graph batch from G_pre do
        /* Step 1: Modality Absence Simulation                       */
        sample absence and reconstruction masks, then set ã_i^(m) ← a_i^(m)(1 − s_i^(m))(1 − r_i^(m));
        /* Step 2: Self-Semantic Initialization and Neighborhood Completion */
        initialize h_i^(0) from visible modalities via Eq. (2);
        compute t_i ← f_topo(p_i) and neighbor reliability r_ij via Eq. (3);
        aggregate incoming neighbors and obtain s_i via Eq. (4);
        /* Step 3: Topology-Guided Shared Latent Calibration         */
        project u_i ← P_sh(s_i) and calibrate z_i^anc via Eq. (5);
        select reliable neighbors R(i) and compute latent constraints via Eq. (6);
        optimize L_pre with original modality targets via Eq. (32);
    end
end
/* Step 4: Unified Downstream Adaptation                             */
initialize the task model from f_Θ and its task interface;
for each task training batch do
    encode task-visible modalities and graph structure to obtain z_i^anc;
    if s is node classification or link prediction then
        select prompt context via Eq. (7) and apply the residual prompt update;
        predict with the downstream head;
    else
        gather the target node and one-hop neighbors into Z_topo;
        map Z_topo → Z̄_s via Eq. (8);
        supply Z̄_s as language prefixes or diffusion context tokens;
    end
    optimize the task objective following Appendix A.6;
end
return pretrained encoder f_Θ and adapted task model;
```

Grocery, Movies, and Toys. These datasets are extracted from the Amazon product review corpus Ni et al. (2019). Nodes represent products, and edges encode purchase relations derived from the also bought metadata. Textual attributes include product titles, descriptions, and reviews, while visual attributes are product images. Grocery and Toys are used for node classification, and Movies is used for few shot node classification. Labels are given by product categories or genres.

RedditS. RedditS is derived from the RedCaps collection of Reddit posts with images and text Desai et al. (2021). We follow the OpenMAG preprocessing protocol to construct its post graph Wan et al. (2026). Post titles and bodies provide textual attributes, and embedded images provide visual attributes. We use RedditS for few shot node classification, where labels identify subreddit communities.

DY and Bili_Dance. These datasets are drawn from the NineRec benchmark Zhang et al. (2024). Nodes represent short videos from Douyin and Bilibili, respectively, and edges reflect user co-viewing or co-interaction behavior. Textual attributes are obtained from titles, descriptions, or hashtags, and visual attributes from video covers or sampled frames. Both datasets are used for link prediction.

Ele-fashion. Ele-fashion is a heterogeneous Amazon graph that merges electronics and fashion products Ni et al. (2019); Hou et al. (2026). Nodes represent products from the two categories, and edges encode cross-category co-purchases. Textual attributes combine technical specifications with style descriptions, while visual attributes are product images. We use Ele-fashion for cross-domain pretraining.

Sports. Sports is drawn from Amazon's Sports and Outdoors category Ni et al. (2019); Hou et al. (2026). Nodes represent sports equipment and accessories, and edges capture complementary or co-purchase relations. Textual attributes comprise product specifications, and visual attributes are product images. We use Sports for cross-domain pretraining.

Flickr30k. Flickr30k Entities provides paired images and sentences with grounding annotations between regions and phrases Plummer et al. (2015). Following the OpenMAG construction Wan et al.

(2026), we link image regions and caption phrases through their grounding relations and evaluate G2Text generation.

SemArt. SemArt is a fine-art collection introduced for multimodal retrieval and semantic art understanding Garcia & Vogiatzis (2018). Following the OpenMAG construction Wan et al. (2026), nodes represent paintings and edges connect artworks sharing metadata such as artist, period, or school. Textual attributes comprise titles and expert descriptions, while visual attributes are the corresponding painting images. We use SemArt for G2Image generation.

A.5 Compared Baselines

We compare GraphBind with representative methods from MAG models, GFMs, and MGFMs. Mario Sun et al. (2026) focuses on adapting LLMs through modality-adaptive graph instruction tuning, with downstream evaluation centered on node classification and link prediction. We leave Mario outside the experimental comparison given its emphasis on LLM-based graph reasoning, while our study mainly focuses on learning shared graph representations and adapting them to both discriminative and generative tasks.

MMGCN Wei et al. (2019) is a multimodal graph convolution model for micro video recommendation. It builds a separate user and item graph for each modality, propagates information independently within each graph, and combines the resulting user and item representations with intrinsic features and shared ID embeddings for personalized ranking.

MGAT Tao et al. (2020) is a multimodal graph attention model for recommendation. It applies gated attention on parallel modality-specific interaction graphs, learns adaptive weights for neighboring items and modalities, and fuses the resulting signals to capture fine-grained user preferences under noisy or conflicting modalities.

LGMRec Guo et al. (2024) combines local graph learning with a global hypergraph module to model collaborative and multimodal user interests for recommendation. Its local branch separates collaborative and modality-specific signals, while the hypergraph branch captures higher-order dependencies among users and items before the two representations are integrated.

TMTE Zhu et al. (2026) jointly and iteratively evolves graph topology and multimodal representations under downstream task supervision, forming a task-aware co-evolution framework for multimodal graph learning. It learns multi-perspective similarity relations to refine the adjacency structure and uses the evolved topology for smoothness-regularized multimodal fusion and cross-modal alignment.

RoleMAG Zuo et al. (2026) routes neighbors according to learned roles and propagates multimodal information through role-specific experts, targeting robust representation learning on multimodal graphs. Its router decomposes each edge into shared, complementary, and heterophilous relations, and the corresponding experts perform smoothing, cross-modal completion, and signed structural filtering before gated fusion.

GIT Wang et al. (2025b) represents graph computation patterns as task trees and pretrains a shared graph model through task-tree reconstruction. A task-tree representation unifies node-, edge-, and graph-level objectives, after which the pretrained encoder supports supervised, few-shot, zero-shot, and in-context adaptation across graph domains.

GFT Wang et al. (2024) learns a transferable vocabulary of computation trees with vector quantization and uses it to unify graph pretraining across domains and downstream fine tuning. During pretraining, graph computation trees are reconstructed into a shared codebook. During adaptation, the resulting tree tokens provide a common interface for node, link, and graph prediction.

OpenGraph Xia et al. (2024) distills graph generalization from large language models through a unified graph tokenizer, a scalable graph Transformer, and LLM-assisted graph data augmentation. The tokenizer converts heterogeneous graph neighborhoods into transferable tokens, while the Transformer models global node dependencies for zero-shot transfer across unseen graph domains and tasks.

RiemannGFM Sun et al. (2025) formulates graph representation learning in a Riemannian geometric space and pretrains a transferable graph foundation model for cross-domain node and link

prediction. Its geometric encoder and task-agnostic pretraining objective are designed to accommodate graphs with diverse structural properties while retaining transferable relational geometry.

UniGraph2 He et al. (2025) aligns text and image features with graph structure through frozen modality encoders, a mixture-of-experts alignment module, and a universal GNN. The experts select modality-specific alignment functions before topology-aware message passing, producing a unified embedding space that transfers to node classification and link prediction.

PLANET Liu et al. (2026) is a multimodal graph foundation model based on a divide-and-conquer design that separately models graph structure and modality semantics before integrating them for transfer across discriminative and generative tasks. Its modular decomposition isolates graph encoding, multimodal representation learning, and task adaptation, enabling a shared foundation model to support heterogeneous downstream objectives.

A.6 Experiment Setup

Experimental Environment. All experiments are implemented in Python 3.10.18 with PyTorch 2.8 and CUDA 12.9. We use a workstation equipped with Intel Xeon Scalable processors, 256 GB system memory, and NVIDIA RTX 6000 Ada GPUs with 48 GB VRAM each.

Pretraining Data and Sampling. We pretrain on Toys, Grocery, Movies, RedditS, Ele-fashion, and Sports. Each dataset contributes one graph update per epoch, giving every dataset one sixth of the updates. Ele-fashion and Sports use sampled subgraphs to control memory use. Textual and visual node attributes are encoded as 768 dimensional vectors by CLIP ViT L/14 Radford et al. (2021) and used as the two modality inputs. DY, Bili_Dance, Flickr30k, and SemArt are reserved for adaptation and evaluation. GraphBind uses a 768 dimensional shared space, two topology GNN layers, a 32 dimensional topology input, eight attention heads, and an eight token topology prompt. Independent absence uses a 10% missing probability for each modality. Neighborhood correlated absence masks one hop neighborhoods of seed nodes with a 3% seed ratio. Property dependent absence scales the missing probability by node properties and sets the average missing ratio to 10% Wu et al. (2026). Modalities visible to the encoder are additionally masked at 15% to provide reconstruction targets.

Pretraining Optimization. We optimize all GraphBind parameters with AdamW Loshchilov & Hutter (2019) for at most 500 epochs using learning rate 1×10^{-4}, weight decay 1×10^{-5}, gradient clipping at norm 1.0, and cosine decay to 1×10^{-6} with 20 warmup epochs. Early stopping is enabled after epoch 300 with patience 150 and minimum improvement 10^{-4}. The InfoNCE temperature is 0.07 and its chunk size is 512. Reliable neighbor completion uses two layers, top 4 latent consistency neighbors, and a reliability threshold of 0.2. The pretraining objective combines contrastive and generative losses with unit weights, decorrelation weight 0.05, latent reconstruction weight 0.3, and latent consistency weight 0.2. The cross modal reconstruction coefficient is 0.5, and hard negative mining considers at most 4,096 nodes with top 4 hard negatives.

Discriminative Fine Tuning. For node classification, we use a 0.6/0.2/0.2 train/validation/test split generated by a seeded random permutation. We train for 100 epochs with AdamW Loshchilov & Hutter (2019) using learning rate 10^{-3} and weight decay 10^{-5}, and stop after 30 epochs without validation improvement. The topology prompt has dropout 0.5. The text classifier is selected with fixed fusion weight 1.0. Few shot node classification uses $K \in \{3, 5\}$ labeled nodes per class, 10 episodes, and the same prompt interface. For link prediction, we train for 200 epochs with learning rate 10^{-3}, weight decay 10^{-5}, and patience 50. We use 0.1/0.2 validation test ratios, one training negative and 1,000 evaluation negatives per positive edge, edge and score batches of 2,048, two hop neighbor fanouts of 15 and 15, and a three layer MLP predictor with hidden size 256 and dropout 0.02. The encoder learning rate scale is 0.1, and the gradient clipping threshold is 1.0. Our implementation computes Hits@1, Hits@3, and Hits@10; we report MRR and Hits@3.

Generative Fine Tuning. For G2Text, we follow the OpenMAG task formulation Wan et al. (2026). All methods use OPT-125M Zhang et al. (2022) with LoRA Hu et al. (2022) of rank 64 and scaling 128 applied to `q_proj` and `v_proj`. GraphBind constructs graph prefixes with the G2M-Adapter, while the compared baselines use a shared prefix projection module. We train for 15 epochs with batch size 4 per device, learning rate 10^{-4}, weight decay 10^{-2}, AdamW Loshchilov & Hutter (2019) parameters $(\beta_1, \beta_2) = (0.9, 0.95)$, gradient clipping 1.0, maximum input and output lengths 96 and 64, and beam search with four beams with maximum generation length 64. The graph backend has

hidden size 768, two layers, eight heads, dropout 0.1, and at most eight neighbor slots. For G2Image, we follow the OpenMAG G2Image protocol Wan et al. (2026) and provide graph tokens to Stable Diffusion v1.5 Rombach et al. (2022). We train at resolution 256×256 for 10 epochs with batch size 16, learning rate 5×10^{-5}, weight decay 10^{-2}, gradient clipping 1.0, and graph context dropout 0.1. Inference uses 50 diffusion steps, guidance scale 7.5, graph context, and up to 4,277 evaluation samples. All methods finetune the Stable Diffusion v1.5 U-Net while keeping the VAE and text encoder frozen. Generated images are evaluated with CLIP ViT-L/14 and DINOv2-large Oquab et al. (2024).

Baseline Configuration. The eleven baselines comprise five MAG models given by MMGCN, MGAT, LGMRec, TMTE, and RoleMAG, four GFMs given by GIT, GFT, OpenGraph, and RiemannGFM, and two MGFMs given by UniGraph2 and PLANET. For the MAG models, hidden dimensions are 128 for MMGCN, MGAT, and LGMRec and 256 for RoleMAG, with their standard layer counts and dropout values. TMTE uses hidden size 64, ten evolution iterations, four similarity perspectives, graph skip coefficient 0.8, and topology update ratio 0.1. RoleMAG uses three shared and backbone layers, modality weights $\lambda_v = \lambda_t = 0.5$, and top 16 complementary neighbors. GIT and GFT use two layer GraphSAGE backbones with hidden size 256. GFT additionally uses a 256 entry codebook. OpenGraph uses a two layer GNN and a two layer graph Transformer with 256 dimensional tokens and 64 anchors. RiemannGFM uses 128 dimensional hidden states and 64 dimensional embeddings. UniGraph2 uses a 768 dimensional three layer encoder with eight experts and top 2 routing. PLANET uses 768 dimensional four layer Transformer blocks, five experts with top 2 routing, a 20,480 entry codebook, and 10 dimensional Laplacian positional encodings. Pretraining checkpoints and task specific optimization settings are selected consistently across methods.

Node Classification. We predict the class of each node on Grocery and Toys. The graph is split into 60%/20%/20% train, validation, and test nodes. A representation from the topology prompt is fed to a linear text classifier trained with cross entropy for 100 epochs using learning rate 10^{-3}, weight decay 10^{-5}, and patience 30. Accuracy is computed as

$$\mathrm{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i), \tag{33}$$

and macro-F1 as

$$\mathrm{F1} = \frac{1}{C} \sum_{c=1}^{C} \frac{2P_c R_c}{P_c + R_c}, \qquad P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c}. \tag{34}$$

Few Shot Node Classification. We evaluate Movies and RedditS with an episodic N way K shot protocol using $N = 5$ and $K \in \{3, 5\}$. In each of 10 episodes, five eligible classes are sampled. Exactly K support nodes are sampled without replacement from the training partition for every selected class, while all validation and test nodes belonging to the selected classes form the corresponding query sets. Labels are remapped to the class index set of the episode. The classifier is selected by validation macro-F1, and accuracy and macro-F1 are averaged across episodes.

Link Prediction. On DY and Bili_Dance, we learn to rank missing destination nodes for each source using one negative edge during training and 1,000 sampled negatives at evaluation. The model is trained for 200 epochs with learning rate 10^{-3}, weight decay 10^{-5}, a three-layer 256-hidden-unit predictor (dropout 0.02), and [15,15] two-hop neighbor sampling. For query q, the positive rank is

$$r_q = \sum_j \mathbb{I}(s_{qj} \geq s_q^+). \tag{35}$$

We report MRR and Hits@K as

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{r_q}, \tag{36}$$

$$\mathrm{Hits@K} = \frac{1}{|Q|} \sum_{q \in Q} \mathbb{I}(r_q \leq K), \qquad K \in \{1, 3, 10\}. \tag{37}$$

G2Text. Following the OpenMAG Flickr30k graph-to-text protocol Wan et al. (2026), the model generates a caption for each center node from its sampled multimodal neighbors. We use the OPT-

125M/LoRA setup above. The main experiments report BLEU-4 Papineni et al. (2002) and ROUGE-L Lin (2004), while CIDEr Vedantam et al. (2015) is additionally provided as a supplementary metric. For candidate corpus c and reference corpus r, BLEU-4 is

$$\text{BLEU-4} = \text{BP} \exp\left(\sum_{n=1}^{4} w_n \log p_n\right), \qquad w_n = \frac{1}{4}, \tag{38}$$

where p_n is the modified n-gram precision and

$$\text{BP} = \begin{cases} 1, & |c| > |r|, \\ \exp(1 - |r|/|c|), & |c| \le |r|. \end{cases} \tag{39}$$

ROUGE-L uses the longest common subsequence length L_{LCS} given below.

$$R_{\text{LCS}} = \frac{L_{\text{LCS}}}{|r|}, \qquad P_{\text{LCS}} = \frac{L_{\text{LCS}}}{|c|}, \tag{40}$$

$$\text{ROUGE-L} = \frac{(1+\beta^2) R_{\text{LCS}} P_{\text{LCS}}}{R_{\text{LCS}} + \beta^2 P_{\text{LCS}}}. \tag{41}$$

CIDEr averages the cosine similarity between TF IDF weighted n gram vectors over $n = 1, \ldots, 4$ and M references.

$$\text{CIDEr}(c, r) = \frac{10}{4M} \sum_{n=1}^{4} \sum_{i=1}^{M} \frac{g_n(c) \cdot g_n(r_i)}{\|g_n(c)\|_2 \|g_n(r_i)\|_2}. \tag{42}$$

G2Image. Following the OpenMAG SemArt graph to image protocol Wan et al. (2026), diffusion guided by graph tokens synthesizes the image associated with each center node. We report CLIP-S and DINO-S. Let $e^{\text{CLIP}}_{\text{gen},i}$ and $e^{\text{CLIP}}_{\text{ref},i}$ denote the ℓ_2 normalized CLIP ViT-L/14 image features Radford et al. (2021). CLIP S is computed as

$$\text{CLIP-S} = \frac{1}{N} \sum_{i=1}^{N} \max\left(\left\langle e^{\text{CLIP}}_{\text{gen},i}, e^{\text{CLIP}}_{\text{ref},i} \right\rangle, 0\right). \tag{43}$$

Similarly, with ℓ_2-normalized DINOv2 features Oquab et al. (2024), DINO-S is

$$\text{DINO-S} = \frac{1}{N} \sum_{i=1}^{N} \max\left(\left\langle e^{\text{DINO}}_{\text{gen},i}, e^{\text{DINO}}_{\text{ref},i} \right\rangle, 0\right). \tag{44}$$

Percentage-style tables multiply these means by 100.