MOBI: Monolithic Graph-Language Modeling Beyond Modality Interference

Zhiyao Zhou
College of Computer Science and Technology, Zhejiang University
Hangzhou, Zhejiang, China
zjucszzy@zju.edu.cn

Yugang Ji
Alibaba Group
Hangzhou, Zhejiang, China
yugang.jyg@alibaba-inc.com

Ziwen Xu
College of Computer Science and Technology, Tongji University
Shanghai, China
2253886@tongji.edu.cn

Zhuonan Zheng
Zhejiang Key Lab of Accessible Perception and Intelligent Systems, Zhejiang University
Hangzhou, Zhejiang, China
zhengzn@zju.edu.cn

Sheng Zhou*
Zhejiang Key Lab of Accessible Perception and Intelligent Systems, Zhejiang University
Hangzhou, Zhejiang, China
zhousheng_zju@zju.edu.cn

Weigao Wen
Alibaba Group
Hangzhou, Zhejiang, China
weigao.wwg@alibaba-inc.com

Jiawei Chen
College of Computer Science and Technology, Zhejiang University
Hangzhou, Zhejiang, China
sleepyhunt@zju.edu.cn

Ming Gu
Zhejiang Key Lab of Accessible Perception and Intelligent Systems, Zhejiang University
Hangzhou, Zhejiang, China
gmwork@zju.edu.cn

Chun Chen
College of Computer Science and Technology, Zhejiang University
Hangzhou, Zhejiang, China
chenc@zju.edu.cn

Can Wang
College of Computer Science and Technology, Zhejiang University
Hangzhou, Zhejiang, China
wcan@zju.edu.cn

Abstract

Graph-Language Models (GLMs) aim to endow LLMs with structure-grounded reasoning ability, yet existing solutions often struggle with *modality interference*: structural information can disrupt pretrained linguistic reasoning, while language cues can overwhelm structural signals. Mainstream modular GLMs with an external graph encoder attempt to mitigate the interference by separating graph encoding from language decoding. This separation fails to strike an effective balance between modality fusion and interference, exhibiting limited cross-modal interaction while leaving interference between modalities largely unresolved. To tackle the above challenges, we propose MOBI (**Mo**nolithic Graph-Language Modeling Beyond **M**odality **I**nterference), a monolithic graph-language model that unifies graph encoding and language decoding within a single backbone for end-to-end graph-text fusion. Specifically, MOBI overcomes modality interference via (i) dual-pathway transformer that preserves pretrained linguistic knowledge while acquiring structural understanding, (ii) progressive interaction scheduling that suppresses cross-modal noise by dynamically regulating the information flow, and (iii) correlation-guided attribute perturbation that discourages textual shortcuts on text-attributed graphs. Extensive experiments on 11 datasets across various settings demonstrate that MOBI consistently outperforms modular baselines.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**.

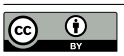
Keywords

Graph-Language Model, Modality Interference

ACM Reference Format:
Zhiyao Zhou, Yugang Ji, Ziwen Xu, Zhuonan Zheng, Sheng Zhou, Weigao Wen, Jiawei Chen, Ming Gu, Chun Chen, and Can Wang. 2026. MOBI: Monolithic Graph-Language Modeling Beyond Modality Interference. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3818175

Resource Availability:
The source code of this paper has been made publicly available at https://doi.org/10.5281/zenodo.20429953.

*Sheng Zhou is the corresponding author.



KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818175

1 Introduction

Traditional graph learning models, such as Graph Neural Networks (GNNs) [17, 26, 45] and Graph Transformers [53, 54], have achieved strong performance on curated benchmarks; however, they often fall short in real-world settings where labels are scarce or unavailable, thereby heavily relying on task-specific supervision [21, 59]. Recent breakthroughs in Large Language Models (LLMs) [2, 42, 44] have revealed a compelling route toward general-purpose, zero-shot reasoning. Notably, LLMs have been successfully extended beyond text to other modalities, giving rise to powerful multimodal LLMs, such as Vision-Language Models (VLMs) [1, 30, 34], as well as audio-language [11, 23] and video-language models [32, 38] that jointly integrate perception signals and language understanding. These advances naturally motivate treating graphs as another modality, spurring growing interest in Graph-Language Models (GLMs) [3, 27, 41] that seek to bridge graph-structured data with LLMs.

Graphs and language are inherently different modalities: graphs represent relational inductive biases through permutation-invariant topology, while language organizes semantics as an ordered sequence. Jointly modeling them therefore faces a core problem of *modality interference*, where learning structural perception can negatively affect the LLM's linguistic capabilities, and vice versa.

To mitigate this problem, most existing GLMs follow an "Align-and-Connect" paradigm [4, 41, 47, 50], mirroring the prevalent *modular* design in VLM literature [30, 34]. Concretely, an external graph encoder (typically a GNN) produces graph tokens that are passed by a lightweight connector (e.g., a shallow projector) to a (largely) frozen LLM, aiming to reduce modality interference by isolating graph encoding from language decoding. However, this isolation comes at the cost of limited graph-text fusion: the graph encoder must compress topology into a narrow interface and struggles to generalize across diverse graph distributions [21, 33], while the LLM cannot directly access raw graph topology and thus remains bounded in structure-grounded reasoning [55]. To further strengthen graph-text interaction, several recent works [27, 48] introduce heavyweight modules to jointly encode two modalities. However, they remain confined to the modular paradigm. The heavyweight encoder also incurs substantial computational overhead and inference latency [8]. Overall, these methods fail to strike an effective balance between modality fusion and interference, exhibiting limited cross-modal interaction while leaving interference between modalities largely unresolved.

To overcome the aforementioned limitations, we explore moving beyond the modular framework, achieving deep graph-text integration while effectively mitigating modality interference. Motivated by the success of emerging encoder-free architectures in VLM research [9, 28, 35], we aim to develop a monolithic graph-language model with native structural understanding for generalizable graph learning. Unlike prevalent modular approaches, a *monolithic* framework employs a single Transformer to unify graph encoding and language decoding, enabling end-to-end learning of deep, flexible graph-text interactions.

Nevertheless, developing a monolithic GLM to go beyond modality interference is non-trivial, which translates into the following challenges, as shown in Figure 1: *i)* Relying on lightweight projector

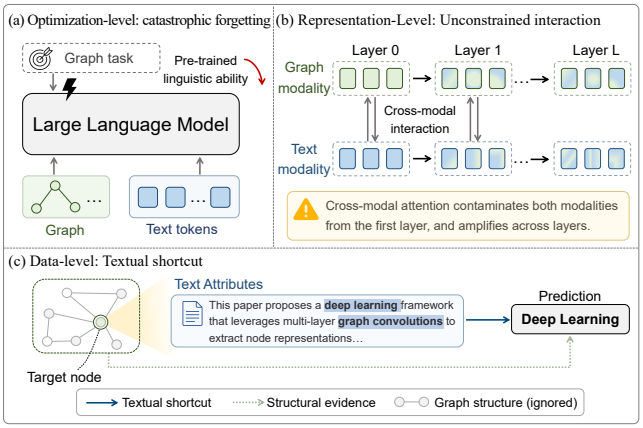


Figure 1: Three levels of modality interference in monolithic graph-language models.

tuning is inadequate for capturing generalizable structural knowledge [48, 55], whereas fully fine-tuning the LLM risks catastrophic forgetting of pre-trained linguistic knowledge (we empirically find that directly fine-tuning the model causes the invalid output rate to surge in downstream tasks). This optimization-level interference, also known as the *stability-plasticity dilemma*, necessitates a specialized model capable of learning new structural patterns while shielding the linguistic knowledge from interference. *ii)* While a monolithic architecture enables free graph-text interplay, unconstrained attention in shallow layers can entangle immature representations and propagate cross-modal noise. We diagnose this representation-level interference by tracking cross-modal attention weights across layers: without explicit regulation, cross-modal attention dominates from layer 1 (>80%) for graph tokens, whereas a well-regulated model should exhibit near-zero cross-modal attention in early layers. This requires a mechanism to dynamically regulate the information flow. *iii)* Given the rich text attributes in common TAGs [10, 51], LLMs are prone to establishing textual shortcuts while ignoring topological signals during training [22, 49]. This data-level interference results in performance degradation in scenarios where structural information dominates.

To address these challenges, we propose MOBI (**Mo**nolithic Graph-Language Modeling **B**eyond Modality **I**nterference), a novel monolithic framework designed for robust zero-shot graph generalization. Unlike modular baselines, MOBI operates without external encoders, directly taking graph nodes and structures as input for end-to-end learning. Concretely, MOBI resolves modality interference through three key innovations: *i) The Dual-Pathway Transformer* mitigates catastrophic forgetting by disentangling graph and text parameter pathways, enabling effective structural learning without compromising pre-trained linguistic capabilities; *ii) The Progressive Interaction Scheduling* reduces cross-modal noise by regulating the attention mechanism, prioritizing uni-modal encoding in shallow layers and progressively intensifying graph-text interaction in deeper layers where representations are semantically

aligned; *iii) The Correlation-Guided Attribute Perturbation* selectively masks the textual attributes of nodes exhibiting high text-label correlations, compelling the model to leverage topological information. Collectively, these designs yield a monolithic graph-language model that can perform zero-shot, structure-grounded reasoning. Extensive experiments across 11 datasets show that MOBI consistently outperforms modular baselines.

In summary, our contributions are as follows:

- We propose MOBI, a monolithic graph-language model that enables end-to-end learning of graph-text interactions for structure-grounded reasoning within a single Transformer.
- We introduce three key designs to resolve modality interference, including a Dual-Pathway Transformer to prevent catastrophic forgetting, Progressive Interaction Scheduling to reduce cross-modal noise, and Correlation-Guided Attribute Perturbation to mitigate textual shortcuts. These designs jointly empower the model with native and generalizable structural understanding.
- Extensive experiments demonstrate the superiority of MOBI in zero-shot scenarios across diverse datasets, tasks, and domains. Furthermore, in-depth analyses validate the effectiveness of each component, as well as MOBI's versatility and efficiency.

2 Related Work

2.1 Graph Neural Networks

Graph Neural Networks (GNNs) have become the dominant approach for representation learning on graph-structured data [26, 45, 56]. Extensive studies have developed advanced GNN architectures and achieved strong performance on various tasks [7, 15, 45, 57]. However, their success often relies on abundant high-quality labels, which are costly to obtain in practice. To reduce label dependence, self-supervised graph learning methods leverage contrastive or generative objectives to learn from unlabeled data [43, 46, 59]. Meanwhile, graph prompt-based approaches adapt prompts from NLP to unify different graph tasks under a shared pre-training framework [12, 39]. Despite alleviating label scarcity, most methods still follow the pre-train–fine-tune pipeline and thus require labeled data for downstream adaptation, limiting their zero-shot generalization.

2.2 LLMs for Graph Learning

Existing studies that incorporate Large Language Models (LLMs) into graph learning can be broadly grouped into two categories: LLM as Enhancer and LLM as Predictor [25].

LLM as Enhancer methods exploit LLMs to enrich node features [6, 18], refine graph structures [16, 62], or generate pseudo-labels [60], achieving strong supervised performance by leveraging pre-trained knowledge. However, since these methods primarily utilize GNNs as the final predictive backbone, they still rely on labeled data and thus generalize poorly to zero-shot settings. Recent attempts to improve cross-dataset transfer with language semantics [31, 33] yield only limited gains, mainly due to shallow graph–text interaction and the use of relatively small LLMs.

LLM as Predictor methods primarily follow two distinct paths. The first involves linearizing graphs into text sequences for direct LLM input [13, 49]. However, these methods suffer from prompt sensitivity and excessive sequence lengths [40]. The second path aims to develop Graph-Language Models (GLMs), akin to Vision-Language Models (VLMs) research [1, 30, 34]. Most existing GLMs adopt GNN-LLM modular architectures, where the GNN typically undergoes a CLIP-style [37] alignment process to generate inputs compatible with the LLM [4, 41, 47, 50, 61]. While decoupling graph encoding from language decoding alleviates modality interference, it hinders end-to-end modality fusion. On the one hand, since the LLM is restricted to the graph encoder's lossy compression without accessing raw topology, its structural understanding is inevitably upper-bounded. On the other hand, the lack of task-specific guidance makes training of a universally generalizable inherently challenging.

More recently, some attempts have been made to enhance graph-text fusion. For example, GOFA [27] couples a GNN into an additional LLM as a graph encoder to generate prompt-guided representations from raw text. UniGTE [48] proposes a graph-text encoding module featuring structure-aware graph-text attention. However, these methods focus solely on refining the encoder part and remain confined to the modular paradigm with separate graph encoding and language decoding, while leaving the modality interference problem unsolved. Besides, incorporating a heavyweight encoder incurs substantial computational overhead and inference latency, posing engineering complexities during deployment [8].

Two exceptions of the modular paradigm are LLaGA [3] and GDL4LLM [63] that directly input node features into LLMs based on tree templates or random walk sequences. While removing separate encoders, they fail to fully preserve fine-grained graph structural properties (e.g., permutation invariance). Furthermore, they treat graph tokens and text tokens uniformly without modality-specific specialization and only consider supervised scenario.

Different from all these methods, we propose a monolithic GLM that directly address the modality interference problem. To the best of our knowledge, we are the first to investigate monolithic GLM with native graph structural understanding for zero-shot generalization on graphs.

2.3 Monolithic VLMs

Unlike the prevailing modular paradigm that relies on external vision encoders, monolithic VLMs [8, 9, 28, 29, 35] integrate visual encoding and language decoding into a unified transformer architecture to bypass the inductive biases of pretrained visual encoders, enable end-to-end modality fusion and improve deployment efficiency. EVE [8] pioneered this direction by directly projecting raw image patches into the LLM's token space, bypassing heavy visual encoders. To address the catastrophic forgetting problem in training monolithic VLMs, Mono-InternVL [35] incorporates endogenous visual experts via a Mixture-of-Experts (MoE) architecture to disentangle visual learning from linguistic knowledge. Recently, SAIL [28] has empirically validated the scalability of these simplified, encoder-free architectures, suggesting that a single transformer is sufficient for learning robust multi-modal representations given sufficient data. These studies collectively validate the potential of encoder-free, end-to-end monolithic modeling of multi-modal data. Inspired by these works, we aim to develop a monolithic LGLM for zero-shot generalization on graphs.

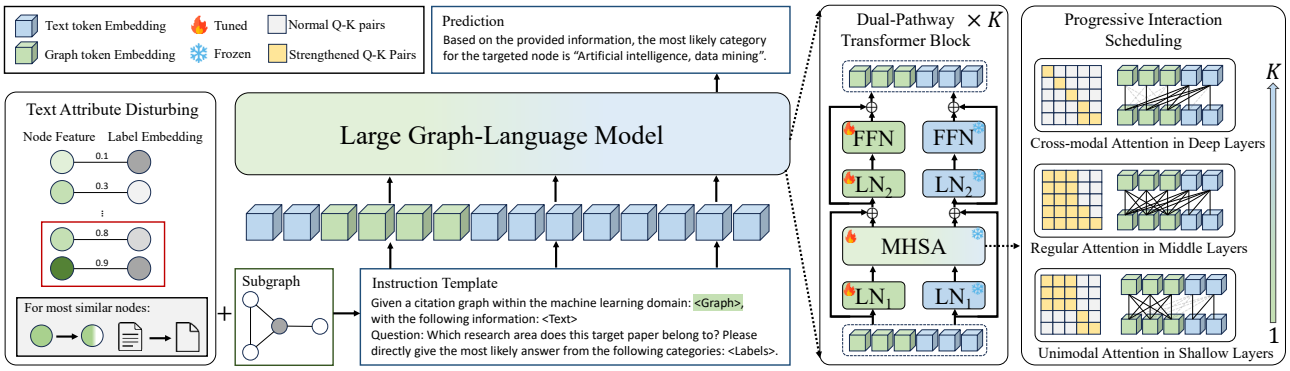


Figure 2: MOBI architecture. The model integrates graph and text processing in a single Transformer with three key mechanisms: correlation-guided attribute perturbation, dual-pathway transformer, and progressive interaction scheduling.

3 Preliminary

Graph is a data structure that describes a set of entities and their relationships. Formally, a graph can be represented as $G = (\mathcal{V}, \mathcal{E}, \mathbf{A}, \mathbf{X})$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ is the set of N nodes, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges. $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the adjacency matrix representing the topological structure, with $A_{ij} = 1$ if an edge exists between node v_i and v_j, and $A_{ij} = 0$ otherwise. $\mathbf{X} \in \mathbb{R}^{N \times F}$ is the node feature matrix, where each row represents the F-dimensional feature vector of a node. These features are typically derived by encoding the raw features of nodes using a Pre-trained Language Model (PLM).

In existing GLM studies, graph tasks are reformulated as a text generation task by converting the graph information and task requirement into a graph-text hybrid input, which is fed into the LLM. Take node classification as an example, a typical graph-text hybrid input [41] can be structured as: *Given a citation graph from arXiv: <node 1><node 2>…<node n>, with the following information: Title: {title}, Abstract: {abstract}. Question: Which arXiv CS sub-category does this target paper belong to?* Here, <node 1><node 2>…<node n> represents a sequence of graph token embeddings. These embeddings contain information of the subgraph G_{sub} centered around the target node (or edge) for node-level (or edge-level) tasks. In the modular paradigm, they are generated by an external graph encoder. In contrast, in our monolithic model, we directly use node features as graph tokens. The graph-text hybrid input sequence can be denoted as $\mathbf{x} = (x_1, x_2, \ldots, x_L)$. To distinguish between modalities, we introduce a corresponding sequence $\mathbf{u} = (u_1, u_2, \ldots, u_L)$, where $u_i \in \{g, t\}$ indicates whether the i-th token belongs to the graph (g) or text (t) modality. Given this hybrid input sequence, the model needs to simultaneously process both graph and language information to generate the final answer.

4 Methodology

We introduce **MOBI**, a monolithic graph-language model capable of structure-aware reasoning for generalizable graph learning. Figure 2 outlines the key components of our proposed method. We employ a **Dual-Pathway Transformer** to endow the LLM with graph structural awareness without compromising its inherent linguistic knowledge. A **Progressive Interaction Scheduling** mechanism is incorporated to dynamically regulate the interaction between graph and text tokens. To mitigate the textual shortcuts, we introduce a **Correlation-Guided Attribute Perturbation** strategy during training. This strategy encourages the model to prioritize structural patterns over textual cues. The structure of this section is organized as follows: we start by introducing a generalized version of the attention mechanism that supports hybrid graph-text inputs, followed by a detailed presentation of the three key designs we propose.

4.1 Graph-Text Attention

Here we introduce a generalized attention mechanism as the premise of our model. In our monolithic paradigm, a natural question is how to conduct self-attention given the graph-text hybrid input. Most existing approaches [3, 41] simply treat the graph tokens as text tokens and adopt conventional causal attention. We adopt a generalized graph-text attention mechanism [48] that assigns shared positional ids to graph tokens and integrates graph relative positional encoding to perceive the graph structure, as in Graph Transformer models [53, 54]. Specifically, the attention score before applying the softmax function can be formally expressed as:

$$S_{ij} = x_q^T W_Q^T \mathcal{R}(p(i) - p(j)) W_K x_k / \sqrt{d} + b_g(i, j) + b_m(p(i), p(j)), \tag{1}$$

where $p(\cdot)$ is a position mapping function that assigns a shared position ID to all graph tokens while preserving the sequential positions for text tokens. The first term operates as in standard ROPE. $b_g(i, j)$ is a graph relative positional encoding bias that encodes the structural relationship between nodes (e.g., shortest path distance). $b_m(i, j)$ is the causal attention masking bias. This design ensures that graph tokens can attend to each other with structural information injected, while their ordering has no effect on the attention process. However, the generalized graph-text attention alone is insufficient to build a monolithic GLM with native structural understanding due to severe modality interference. Specifically, i) it does not address the stability-plasticity problem. ii) It lacks explicit regulation of graph–text interactions, making it prone to shallow-layer entanglement that undermines effective alignment and fusion. iii) It remains susceptible to spurious text–label correlations on text-attributed graphs.

4.2 Dual-Pathway Transformer

Inspired by the divide-and-conquer principle [9], we employ a hybrid architecture that integrates modality-aware parameter sparsity and cross-modal parameter sharing within a unified Transformer model. Specifically, in the first K *dual-pathway Transformer* blocks of the LLM, we perform parameter decoupling and introduce two pathways, where tokens from different modalities are assigned distinct parameters. In the other *single-pathway Transformer* blocks, parameters are shared between the graph and text modalities. In the dual-pathway Transformer blocks, rather than simply applying the Mixture-of-Experts(MoE) [14] designs to the feed-forward network (FFN), we additionally consider the self-attention and normalization layers to fully capture modality-specific patterns. For the attention layer, we explicitly define separate projection matrices for different token types:

$$\begin{aligned} \mathrm{ATTN}(\mathbf{x};\theta^u_{\text{attn}})_i &= \sum_{j=1}^{L} \frac{\exp(S_{ij})}{\sum_{k=1}^{L}\exp(S_{ik})} W_V^{u_j} x_j, \\ Q_i = W_Q^{u_i} x_i, \quad K_i &= W_K^{u_i} x_i, \quad V_i = W_V^{u_i} x_i. \end{aligned} \tag{2}$$

Here the modality-specific query, key, and value are derived from their respective attention weight matrices $W^{u_i}, u_i \in \{g, t\}$. Note the first score term in Eq (1) becomes $x_q^T (W_Q^{u_i})^T \mathcal{R}(p(i) - p(j)) W_K^{u_j} x_k$ in this case. This enables the LLM to model diverse patterns within both unimodal and cross-modal attention. The dual-pathway Transformer block can be expressed as follows:

$$\begin{aligned} \mathbf{h} &= \mathbf{x} + \mathrm{ATTN}\big(\mathrm{LN1}(\mathbf{x};\theta^u_{\text{ln1}});\theta^u_{\text{attn}}\big), \\ \mathbf{x}' &= \mathbf{h} + \mathrm{FFN}\big(\mathrm{LN2}(h;\theta^u_{\text{ln2}});\theta^u_{\text{ffn}}\big). \end{aligned} \tag{3}$$

During model training, we update the parameters of the graph modality while freezing those of the text modality. This enables the LLM to develop graph-awareness from scratch while simultaneously preserving its pretrained knowledge by freezing the textual parameters, as illustrated in Figure 2. In the first K dual-pathway blocks, distinct parameter sets enable the model to capture modality-specific patterns and align graph and text representations. Then, the later single-pathway blocks conduct structure-grounded reasoning on the aligned representation. It is worth noting that this form of decoupling differs from the separation of graph encoding and language decoding in modular architectures. Despite using distinct parameter sets, the graph and text modalities can still interact within the self-attention layers with structural information injected.

4.3 Progressive Interaction Scheduling

Although the architectural innovation described above ensures that the parameters of the text modality remain unaffected during the learning of graph structural perception, the modality interference persists at the representation level, where unconstrained interaction in the shallow layers can entangle immature representations and amplify cross-modal noise. Intuitively, each modality should focus on aggregating modality-specific information in the shallow layers, transitioning to cross-modal interaction in deeper layers when robust uni-modal representations have been learned.

Therefore, we propose a progressive interaction scheduling mechanism applied to K dual-pathway Transformer blocks. Specifically, we add an additional dynamic modality-wise masking bias term $b_c(u_i, u_j, k)$ added to S_{ij}:

$$\hat{S}_{ij} = S_{ij} + b_c(u_i, u_j, k), \tag{4}$$

where the added bias term returns the corresponding attention bias based on the modality types of the query and key, and the current layer number $k \in [1, K]$. There exist various options for implementing the function b_c. Our default choice is a linear function that decays with layer depth, formulated as:

$$b_c(u_i, u_j, k) = \begin{cases} b_g^{\text{start}} + \frac{k-1}{K-1}(b_g^{\text{end}} - b_g^{\text{start}}) & \text{if } u_i = u_j = g \\ b_t^{\text{start}} + \frac{k-1}{K-1}(b_t^{\text{end}} - b_t^{\text{start}}) & \text{if } u_i = u_j = t \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where b_g^{start}, b_t^{start}, b_g^{end} and b_t^{end} are hyperparameters denoting the initial&final bias values for the first&last layer, respectively. The incorporated attention bias can be interpreted as a regulator of intra-modality information exchange, which gradually decreases as the network goes deeper. Note that we only consider the graph–graph and text–text pairs here, since this is equivalent to considering all four possible modality combinations in the attention scores before softmax. We also consider implementing b_c as a learnable gating function, but found that it brought little improvement (details in Section 5.4).

The proposed modality-wise attention bias term imposes a hierarchical, curriculum-like control over information flow of the model. In the early layers, by focusing on intra-modality attention, the model can develop robust and high-quality uni-modal representations. As the model transitions to deeper layers, it progressively shifts its focus to cross-modal fusion and alignment, effectively addressing the noise propagation that occurs in graph-text interactions. In the final few dual-pathway Transformer layers, the bias term is assigned a negative value so as to prioritize cross-modal interactions over intra-modal ones. This design is intended to enable the K modality-aware Transformer layers to more effectively align graph structures with language tokens. Note that the following standard Transformer layers continue to employ the standard attention mechanism.

Interestingly, with the modality-wise attention bias term and the dual-pathway Transformer blocks, we can regard a modular architecture as a special case of our framework to some extent. When we impose a hard constraint that completely prohibits cross-modal interaction in the first M blocks, the graph-modality component in this part functions as an independent M-layer graph Transformer, whose output is passed to the subsequent Transformer blocks. The corresponding modality-wise attention bias function can be expressed as:

$$b_c(u_i, u_j, k) = \begin{cases} \mu, & \text{if } (u_i = u_j) \wedge (k < M), \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where μ is a large positive value ($\mu \gg |S_{ij}|$). This highlights the advantage of our method from another perspective: the use of soft attention constraints offers greater flexibility and adaptability, allowing a small amount of critical cross-modal information to propagate in the shallow layers (e.g., domain type or task type can guide different patterns of structural perception).

4.4 Correlation-Guided Attribute Perturbation

Another challenge in training a monolithic GLM arises from the inherent bias of existing text-attributed graph (TAG) datasets, where LLMs tend to exploit spurious correlations between text attributes and target labels [10, 49]. When a node's textual description is semantically close to its label (e.g., a paper titled Graph Neural Networks categorized under "Artificial intelligence, neural networks"), the LLM may correctly predict the label by relying solely on text, thereby ignoring topological information from the graph structure. Such reliance on textual shortcuts can undermine the model's generalization ability, particularly in downstream datasets where textual cues are less informative and the graph structure plays a critical role.

To mitigate this issue, we introduce a correlation-guided attribute perturbation strategy during training. Specifically, we first compute the embeddings of label texts using a PLM, and measure the cosine similarity between node feature and its corresponding label embedding for each node. A high similarity score serves as a strong indicator that the node's label can be easily inferred from its text description alone. We then apply dual-pronged perturbation to the top p of nodes with the highest similarity scores. For these selected nodes, (i) we completely remove their text attributes from the input prompts provided to the LLM, and (ii) we randomly mask a subset of dimensions in their numerical features with zeros. This strategy encourages the LLM to rely on the graph's topological signals for accurate prediction on the perturbed samples, thereby reducing its dependence on textual shortcuts and promoting a deeper understanding of structural relationships.

5 Experiments

To validate the effectiveness of our proposed MOBI, we conduct a series of empirical evaluations, guided by the following research questions:

- RQ1: How does MOBI perform on unseen datasets within the same domain (i.e., cross-dataset zero-shot generalization)?
- RQ2: Can MOBI effectively generalize to more challenging cross-domain and cross-task scenarios?
- RQ3: Are all components of MOBI necessary and beneficial to its final performance?
- RQ4: How sensitive is MOBI to different LLM backbones and key hyperparameters?

5.1 Experimental Settings

5.1.1 Datasets. We conduct comprehensive experiments on 11 widely used datasets. These datasets cover various domains, including citation networks (Arxiv [20], Pubmed [18], Cora [52]), e-commerce graphs (Computer [58], Books-children [58], Books-History [58], Photo [58], Sports [58]), web links (WikiCS [36]), and social networks (Instagram [24], Reddit [24]). These datasets exhibit distinct characteristics, ensuring a thorough evaluation of MOBI's efficacy. Following previous work [47], we instruction-tune our model on the node classification task using Arxiv and Computer, respectively, and evaluate its cross-dataset transferability on the corresponding downstream datasets within the same domain. We further assess cross-domain transferability on unseen domains, namely web links and social networks. To examine cross-task generalization, we evaluate the model on the unseen link prediction task. For all datasets, we adopt the same data splits as in [47]. Detailed descriptions and splitting protocols for these datasets can be found in Appendix A.

5.1.2 Baselines. We compare our proposed method with a wide range of cutting-edge baselines, categorized into four major groups: i) Classic GNNs: We include representative supervised graph neural networks, namely GCN [26], GraphSAGE [17], and GAT [45]. ii) Self-supervised Graph Learning: We select DGI [46] as a representative method for self-supervised representation learning. iii) Large Language Models (LLMs): We select Vicuna-7B-v1.5 [5] as a pure text-based baseline, which is also used as our LLM backbone. iv) The latest models equipped with zero-shot capabilities: This group comprises recent methods that employ LLMs/LMs for zero-shot graph tasks, including OFA [33], ZeroG [31], LLaGA [3], GraphGPT [41], TEA-GLM [47], and GOFA [27].

5.1.3 Implementation Details. For node prediction tasks, we employ Accuracy and Macro-F1 as evaluation metrics. For link prediction, we report the Area Under the Curve (AUC). Since standard GNN-based methods are not inherently applicable for zero-shot scenarios, we adopt a linear probing protocol as described in [41]. Specifically, we pre-train the GNN backbone on the source dataset, freeze the parameters, and only re-train a linear classification head on the downstream dataset. For our proposed MOBI, we train the model on the source dataset for 2 epochs. The number of modality-aware transformer layers K is tuned within the set {2, 4, 6, 8, 10}. In the progressive graph-text interaction scheduling mechanism, we adopt the linear version as the default setting, and the initial and final biases are constrained to have equal absolute values, denoted as b and selected from {1, 2, 3, 4}. The ratio of samples to disturb p is selected from {15%, 40%, 80%}. Unless otherwise specified, other hyperparameters regarding model configuration and training are kept at their default values. More details on hyperparameters can be found in Appendix C.

5.2 Cross-Dataset Zero-shot Generalization (RQ1)

Table 1 presents the performance of all baselines and our proposed method on the zero-shot cross-dataset setting. The key observations are as follows:

First, traditional GNN-based methods (i.e., GCN [26], GraphSAGE, GAT, and DGI) fail to generalize to unseen datasets. They exhibit substantial performance degradation in the zero-shot setting, sometimes approaching random guessing. This failure is primarily due to their heavy reliance on dataset-specific feature distributions and structural patterns. This limitation also undermines LLM-as-Predictor approaches that rely on pre-aligned GNN embeddings, highlighting the inherent difficulty of training a universal graph encoder that generalizes well across diverse graph structures.

Second, among the recently proposed LLM-as-Enhancer methods with zero-shot capabilities, OFA retains a GNN-centric architecture and achieves only marginal improvements. ZeroG makes noticeable progress by reformulating node classification as a transferable text

Table 1: Performance comparison in the cross-dataset zero-shot setting. We report accuracy (Acc) and macro-F1 score (Macro-F1). The best results are highlighted in bold and the runner-ups are underlined.

Model	PubMed		Cora-large		Children		History		Photo		Sports	
	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1
GCN	0.288	0.187	0.017	0.007	0.030	0.006	0.063	0.024	0.103	0.034	0.042	0.017
SAGE	0.315	0.257	0.014	0.007	0.008	0.005	0.195	0.029	0.056	0.020	0.051	0.021
GAT	0.343	0.259	0.016	0.006	0.080	0.063	0.172	0.159	0.050	0.036	0.142	0.091
DGI	0.329	0.213	0.020	0.004	0.082	0.012	0.218	0.038	0.224	0.045	0.049	0.018
OFA	0.314	0.287	0.130	0.091	0.064	0.017	0.052	0.026	0.340	0.103	0.101	0.043
ZeroG	0.760	0.743	0.162	0.141	0.176	0.177	0.475	0.206	0.455	0.409	0.389	0.450
Vicuna-7B-v1.5	0.744	0.757	0.163	0.120	0.257	0.271	0.352	0.336	0.505	0.408	0.326	0.330
LLaGA	0.793	0.778	0.168	0.108	0.199	0.163	0.146	0.144	0.276	0.362	0.352	0.446
GraphGPT	0.784	0.740	0.148	0.081	0.249	0.162	0.363	0.415	0.547	0.511	0.224	0.224
TEA-GLM	0.853	0.846	0.188	0.126	0.258	0.238	0.533	0.351	0.512	0.467	0.384	0.366
GOFA	0.768	0.741	0.206	**0.175**	0.206	**0.292**	0.394	0.395	0.410	0.461	0.305	0.355
Ours	**0.897**	**0.869**	**0.224**	0.168	**0.277**	0.274	**0.542**	**0.449**	**0.595**	**0.563**	**0.416**	**0.476**

retrieval task. Nevertheless, the absence of deep graph–text interaction in its late fusion architecture and the use of relatively small LM limit its performance gains on more challenging e-commerce datasets.

Third, while LLM-as-Predictor methods exhibit certain generalization capabilities, their improvements over the vanilla LLM backbone are often marginal. Among them, LLaGA exhibits unsatisfactory performance on e-commerce datasets that have varying distributions [47], indicating that merely fine-tuning a projector is insufficient to capture transferable graph knowledge, which is consistent with the findings reported in [55]. Although GraphGPT, TEA-GLM, and GOFA perform well on specific datasets, their improvements are unstable across different datasets. In several cases, these complex methods even underperform the LLM backbone. These phenomena indicate that these methods fail to endow the LLM with generalizable graph structural understanding.

Finally, across six benchmark datasets with twelve metrics, our method achieves state-of-the-art performance on 10 metrics and second-best on the remaining two, demonstrating robust and consistent superiority over existing approaches. This indicates that our model successfully acquires a generalizable capability to understand graph structures while avoiding catastrophic forgetting and cross-modality interference.

5.3 Cross-domain and Cross-task Zero-shot Generalization (RQ2)

To further evaluate the generalization ability of MOBI, we conduct experiments under more challenging scenarios, including transferring to unseen domains (Web Links and Social Networks) and an unseen task (Link Prediction). The results are presented in Table 2.

A key challenge in cross-domain transfer is handling the variation in the relative importance of textual semantics versus structural topology across different graph domains. On text-rich domains (e.g., WikiCS), the inherent semantic capability of the pure LLM plays a dominant role (e.g., vanilla Vicuna-7B achieving 0.641), while most methods suffer from severe negative transfer. In contrast, MOBI reaches 0.695, effectively augmenting the LLM with structural signals without compromising its semantic pretrained knowledge. On structure-dominated domains (e.g., Instagram), where Vicuna-7B collapses to 0.344, MOBI maintains a robust performance of 0.590, significantly outperforming both the LLM backbone and other baselines. This confirms that MOBI possesses transferable structural understanding that remains effective across domains with either rich or sparse textual semantics.

With regard to link prediction task, most baselines exhibit poor generalization with AUC scores around 0.5. This suggests that their learned representations are overly specialized in node-level semantics and fail to encode the topological connectivity essential for link prediction. Our method achieves an AUC score of 0.746 on Pubmed and 0.640 on Photo, demonstrating that MOBI captures topological properties that are transferable and beneficial across different downstream tasks.

5.4 Ablation Study (RQ3)

To accurately reflect the contribution of each component within our model, we conduct a comprehensive evaluation of various MOBI variants.

5.4.1 Analysis of Dual-Pathway Transformer. MOBI decouples graph and text parameter pathways within the Transformer, updating only the graph-modality parameters during training while keeping the text-modality parameters frozen. To validate the effectiveness of our modality-aware parameter decoupling strategy, we compare it against three standard fine-tuning strategies that do not decouple modality-specific parameters: i) Projector-only: fine-tuning only the projection layer that maps graph tokens into the LLM's token space, as in [3]; ii) Full Fine-tuning: updating all parameters of the LLM backbone without any modality-aware specialization; iii) LoRA+Projector: fine-tuning the projector and applying Low-Rank Adaptation [19] to the LLM. The results are reported in Table 3.

We can observe that fine-tuning only the projector consistently yields suboptimal performance. This confirms that a simple linear mapping lacks the capacity to bridge the substantial semantic

Table 2: Performance comparison in the cross-domain and cross-task zero-shot setting. We report accuracy (Acc) and Area Under the Curve (AUC), respectively. The best results are highlighted in bold and the runner-ups are underlined.

Model	Node Classification (Accuracy)			Link Prediction	
	WikiCS	Instagram	Reddit	PubMed	Photo
OFA	0.362	0.580	0.498	0.481	0.459
Vicuna-7B-v1.5	0.641	0.344	0.316	0.543	0.503
LLaGA	0.601	0.397	0.499	0.569	0.478
GraphGPT	0.478	0.462	0.527	0.502	0.485
TEA-GLM	0.449	0.479	0.491	0.689	0.545
GOFA	0.668	0.546	**0.550**	0.507	0.504
Ours	**0.695**	**0.590**	0.545	**0.746**	**0.640**

Table 3: Ablation study on modality-aware parameter decoupling. The best results are highlighted in bold.

Method	Children	History	Photo	Sports
Projector-only	0.224	0.397	0.492	0.338
Full Fine-tuning	0.218	0.343	0.560	0.349
LoRA + Projector	0.262	0.431	0.557	0.396
Ours	**0.277**	**0.542**	**0.595**	**0.416**

gap between graph structures and natural language, failing to fully exploit the LLM's reasoning potential. Full parameter fine-tuning exhibits significant instability, suggesting that aggressively updating all parameters on limited graph instruction data may lead to overfitting or catastrophic forgetting of the pre-trained general knowledge. LoRA+Projector provides relatively stable improvements due to its ability to relieve overfitting and catastrophic forgetting. However, due to the lack of modality-aware specialization, the LLM remains inevitably affected. In contrast, our modality-aware parameter decoupling achieves the best performance across all datasets. This substantial performance improvement indicates that, by assigning distinct parameter sets to process graph and text modalities separately, our method effectively learns graph structural patterns while preserving the pretrained knowledge.

5.4.2 Analysis of Progressive Interaction Scheduling. In Section 4.3, we have introduced three distinct strategies for modulating the modality-wise attention bias function. Here, we compare the performance of these choices. As visualized in Figure 3, removing the bias term consistently yields the worst results. This indicates that unregulated attention leads to severe cross-modal interference: The model mixes unrefined graph and text representations in shallow layers before clean uni-modal representations are obtained. Applying a hard masking strategy generally outperforms the unconstrained baseline by enforcing uni-modal encoding in the shallow layers. The linear version consistently achieves strong performance across all four datasets. By progressively adjusting the attention bias across layers, this design allows the model to balance uni-modal encoding in the early layers with cross-modal fusion in the deep layers. Finally, the learnable gate function exhibits no clear advantage over the linear counterpart. This may be attributed

Table 4: Ablation study on different perturbation choices. The best results are highlighted in bold.

Method	Children	History	Photo	Sports
w/o Perturbation	0.240	0.492	0.547	0.377
Text Removal Only	0.245	0.512	0.538	0.401
Feature Perturbation Only	0.265	0.512	0.566	0.388
Ours	**0.277**	**0.542**	**0.595**	**0.416**

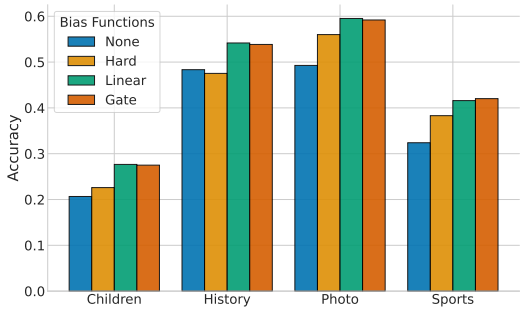


Figure 3: Ablation study on different choices of modality-wise attention bias function.

to the difficulty of optimizing additional parameters with limited graph data, whereas the linear function provides a more effective heuristic prior.

5.4.3 Analysis of Text Attribute Perturbation. To verify the effectiveness of our text attribute perturbation strategy, we consider three variants: i) W/o Perturbation (standard training without textual disturbance); ii) Text Removal Only; iii) Feature Perturbation Only. The experimental results are presented in Table 4. As shown in the table, the variant without perturbation yields the lowest accuracy, suggesting a tendency to overfit specific textual patterns. Moreover, applying single-mode perturbations (either removing text or disturbing features) brings moderate improvements, validating their utility. Finally, we observe that combining both further enhances model performance, effectively mitigating overfitting and improving generalization on graphs.

5.5 Hyperparameter Sensitivity Analysis

In this section, we investigate how the performance of our model varies with respect to three key hyperparameters, including K (the number of dual-pathway transformer layers), b (the absolute magnitude of the initial and final biases in the linear modality-wise attention bias function) and p (the proportion of samples subjected to text attribute perturbation). The corresponding results are presented in Figure 4, Figure 5, and Table 5, respectively.

5.5.1 Impact of K. We investigate the impact of the number of dual-pathway transformer layers K ranging from 2 to 10. We observe that the fluctuation in accuracy is relatively minor within a reasonable range, suggesting that our model is robust to variations in the number of dual-pathway transformer layers.

Table 5: Impact of parameter p.

Dataset	p = 0%	p = 15%	p = 40%	p = 80%
PubMed	0.870	**0.897**	0.855	0.825
History	0.495	**0.542**	0.496	0.514

Table 6: Inference time compared to vanilla LLM and GOFA.

Sec / sample (s)	Pubmed	Cora	Children	History	Photo	Sports
Vicuna-7B-v1.5	2.33	2.20	2.56	1.69	1.90	2.38
GOFA	3.47	2.72	4.09	7.09	4.05	3.27
Ours	2.71	2.41	2.56	1.88	2.06	2.65

5.5.2 Impact of b. We investigate the impact of b, the absolute magnitude of the initial and final biases in the modality-wise attention bias function. Our observations indicate that introducing a non-zero bias consistently outperforms the baseline ($b = 0$). The performance generally peaks at a moderate b before converging to a relatively lower value, which can be considered as a transition from the linear version to the hard version. This confirms that enforcing a structured transition from modality-specific processing in the early layers to cross-modal fusion in the deeper layers is beneficial for capturing effective representations.

5.5.3 Impact of p. We investigate the impact of the proportion of samples subjected to text attribute perturbation p. We observe that introducing a moderate level of perturbation yields the highest accuracy. Increasing p to a large value (e.g., 80%) leads to performance degradation.

5.6 Efficiency Analysis

Here, we compare the inference time of our approach with the vanilla LLM backbone and GOFA. The results are presented in Table 6. GOFA introduces substantial additional computational overhead relative to the LLM backbone. This overhead mainly stems from its encoder–decoder architecture and the intricate interleaving between GNN and Transformer layers. In contrast, our proposed MOBI integrates unimodal encoding and cross-modal interaction into a unified decoder, which facilitates parallel computation. Consequently, compared with the LLM backbone, our model incurs only marginal additional cost, primarily due to the longer input sequences introduced by the appended graph tokens.

6 Conclusion

This paper introduces MOBI, the first monolithic graph-language model that achieves structure-grounded reasoning beyond modality interference. Unlike prevalent modular approaches that rely on external graph encoders, MOBI unifies graph encoding and language decoding within a single Transformer backbone. Through three key technical innovations including Dual-Pathway Transformer, Progressive Interaction Scheduling, and Correlation-Guided Attribute Perturbation, MOBI achieves deep end-to-end fusion while effectively managing modality interference. Extensive experiments across diverse zero-shot settings demonstrate that MOBI consistently outperforms state-of-the-art modular baselines, with ablation studies confirming the effectiveness of each design. Our findings suggest that a unified, encoder-free architecture is viable to empower LLMs with structure-grounded reasoning capabilities. Future work includes extending MOBI to zero-shot graph-level tasks and adapting it to more challenging graph structures such as heterogeneous graphs and dynamic graphs.

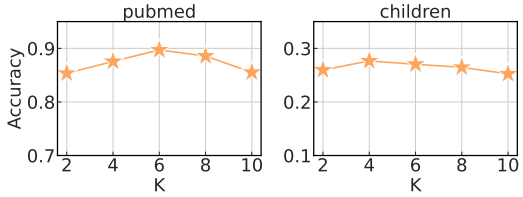


Figure 4: Impact of parameter K.

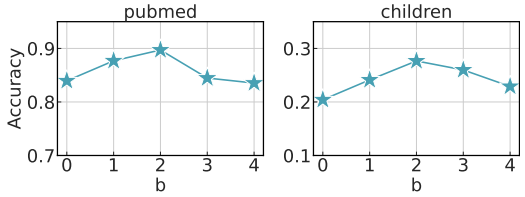


Figure 5: Impact of parameter b.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62476245, No. 62476244), the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study, China (No. SN-ZJU-SIAS-001). This work was also supported by Alibaba Group through Alibaba Innovative Research Program.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Runjin Chen, Tong Zhao, Ajay Kumar Jaiswal, Neil Shah, and Zhangyang Wang. 2024. LLaGA: Large Language and Graph Assistant. In *International Conference on Machine Learning*. PMLR, 7809–7823.

[4] Yongqiang Chen, Quanming Yao, Juzheng Zhang, James Cheng, and Yatao Bian. 2025. Hierarchical Graph Tokenization for Molecule-Language Alignment. In *Forty-second International Conference on Machine Learning*.

[5] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)* 2, 3 (2023), 6.

[6] Eli Chien, Wei-Cheng Chang, Cho-Jui Hsieh, Hsiang-Fu Yu, Jiong Zhang, Olgica Milenkovic, and Inderjit S Dhillon. 2022. Node Feature Extraction by Self-Supervised Multi-scale Neighborhood Prediction. In *International Conference on Learning Representations*.

[7] Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. 2021. Adaptive Universal Generalized PageRank Graph Neural Network. In *International Conference on Learning Representations*.

[8] Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. 2024. Unveiling encoder-free vision-language models. *Advances in Neural Information Processing Systems* 37 (2024), 52545–52567.

[9] Haiwen Diao, Xiaotong Li, Yufeng Cui, Yueze Wang, Haoge Deng, Ting Pan, Wenxuan Wang, Huchuan Lu, and Xinlong Wang. 2025. Evev2: Improved baselines for encoder-free vision-language models. *arXiv preprint arXiv:2502.06788* (2025).

[10] Keyu Duan, Qian Liu, Tat-Seng Chua, Shuicheng Yan, Wei Tsang Ooi, Qizhe Xie, and Junxian He. 2023. Simteg: A frustratingly simple approach improves textual graph learning. *arXiv preprint arXiv:2308.02565* (2023).

[11] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[12] Taoran Fang, Yunchao Zhang, Yang Yang, Chunping Wang, and Lei Chen. 2023. Universal prompt tuning for graph neural networks. *Advances in Neural Information Processing Systems* 36 (2023), 52464–52489.

[13] Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. 2024. Talk like a Graph: Encoding Graphs for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

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

[15] Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. 2019. Predict then Propagate: Graph Neural Networks meet Personalized PageRank. In *International Conference on Learning Representations*.

[16] Zirui Guo, Lianghao Xia, Yanhua Yu, Yuling Wang, Kangkang Lu, Zhiyong Huang, and Chao Huang. 2024. Graphedit: Large language models for graph structure learning. *arXiv preprint arXiv:2402.15183* (2024).

[17] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 1025–1035.

[18] Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. 2024. Harnessing Explanations: LLM-to-LM Interpreter for Enhanced Text-Attributed Graph Representation Learning. In *ICLR*.

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

[20] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems* 33 (2020), 22118–22133.

[21] W Hu, B Liu, J Gomes, M Zitnik, P Liang, V Pande, and J Leskovec. 2020. Strategies For Pre-training Graph Neural Networks. In *International Conference on Learning Representations (ICLR)*.

[22] Jin Huang, Xingjian Zhang, Qiaozhu Mei, and Jiaqi Ma. 2024. Can LLMs Effectively Leverage Graph Structural Information through Prompts in Text-Attributed Graphs, and Why? *Transactions on Machine Learning Research* 2024 (2024).

[23] Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 23802–23804.

[24] Xuanwen Huang, Kaiqiao Han, Yang Yang, Dezheng Bao, Quanjin Tao, Ziwei Chai, and Qi Zhu. 2024. Can gnn be good adapter for llms?. In *Proceedings of the ACM Web Conference 2024*. 893–904.

[25] Bowen Jin, Gang Liu, Chi Han, Meng Jiang, Heng Ji, and Jiawei Han. 2024. Large language models on graphs: A comprehensive survey. *IEEE Transactions on Knowledge and Data Engineering* (2024).

[26] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*.

[27] Lecheng Kong, Jiarui Feng, Hao Liu, Chengsong Huang, Jiaxin Huang, Yixin Chen, and Muhan Zhang. 2025. GOFA: A Generative One-For-All Model for Joint Graph Language Modeling. In *The Thirteenth International Conference on Learning Representations*.

[28] Weixian Lei, Jiacong Wang, Haochen Wang, Xiangtai Li, Jun Hao Liew, Jiashi Feng, and Zilong Huang. 2025. The scalability of simplicity: Empirical analysis of vision-language learning with a single transformer. *arXiv preprint arXiv:2504.10462* (2025).

[29] Han Li, Xinyu Peng, Yaoming Wang, Zelin Peng, Xin Chen, Rongxiang Weng, Jingang Wang, Xunliang Cai, Wenrui Dai, and Hongkai Xiong. 2025. Onecat: Decoder-only auto-regressive model for unified understanding and generation. *arXiv preprint arXiv:2509.03498* (2025).

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[31] Yuhan Li, Peisong Wang, Zhixun Li, Jeffrey Xu Yu, and Jia Li. 2024. Zerog: Investigating cross-dataset zero-shot transferability in graphs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1725–1735.

[32] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*. 5971–5984.

[33] Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. 2024. One For All: Towards Training One Graph Model For All Classification Tasks. In *The Twelfth International Conference on Learning Representations*.

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

[35] Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jiawen Liu, Jifeng Dai, Yu Qiao, and Xizhou Zhu. 2025. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 24960–24971.

[36] Péter Mernyei and Cătălina Cangea. 2020. Wiki-cs: A wikipedia-based benchmark for graph neural networks. *arXiv preprint arXiv:2007.02901* (2020).

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[38] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7464–7473.

[39] Mingchen Sun, Kaixiong Zhou, Xin He, Ying Wang, and Xin Wang. 2022. Gppt: Graph pre-training and prompt tuning to generalize graph neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1717–1727.

[40] Yuanfu Sun, Zhengnan Ma, Yi Fang, Jing Ma, and Qiaoyu Tan. 2025. GraphICL: Unlocking Graph Learning Potential in LLMs through Structured Prompt Design. In *Findings of the Association for Computational Linguistics: NAACL 2025*. 2440–2459.

[41] Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. 2024. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 491–500.

[42] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[43] Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Rémi Munos, Petar Veličković, and Michal Valko. 2021. Bootstrapped representation learning on graphs. In *ICLR 2021 workshop on geometrical and topological representation learning*.

[44] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[45] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

[46] Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. [n. d.]. Deep Graph Infomax. In *International Conference on Learning Representations*.

[47] Duo Wang, Yuan Zuo, Fengzhi Li, and Junjie Wu. 2024. Llms as zero-shot graph learners: Alignment of gnn representations with llm token embeddings. *Advances in Neural Information Processing Systems* 37 (2024), 5950–5973.

[48] Duo Wang, Yuan Zuo, Guangyue Lu, and Junjie Wu. 2025. UniGTE: Unified Graph–Text Encoding for Zero-Shot Generalization across Graph Tasks and Domains. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[49] Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. 2023. Can language models solve graph problems in natural language? *Advances in Neural Information Processing Systems* 36 (2023), 30840–30861.

[50] Runze Wang, Mingqi Yang, and Yanming Shen. 2025. Bridging Molecular Graphs and Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 21234–21242.

[51] Yuxiang Wang, Xinnan Dai, Wenqi Fan, and Yao Ma. 2025. Exploring graph tasks with pure llms: A comprehensive benchmark and investigation. *arXiv preprint arXiv:2502.18771* (2025).

[52] Zhihao Wen and Yuan Fang. 2023. Augmenting low-resource text classification with graph-grounded pre-training and prompting. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 506–516.

[53] Qitian Wu, Chenxiao Yang, Wentao Zhao, Yixuan He, David Wipf, and Junchi Yan. 2023. DIFFormer: Scalable (Graph) Transformers Induced by Energy Constrained

Diffusion. In *The Eleventh International Conference on Learning Representations*.
[54] Qitian Wu, Wentao Zhao, Zenan Li, David P Wipf, and Junchi Yan. 2022. Nodeformer: A scalable graph structure learning transformer for node classification. *Advances in Neural Information Processing Systems* 35 (2022), 27387–27401.
[55] Xixi Wu, Yifei Shen, Fangzhou Ge, Caihua Shan, Yizhu Jiao, Xiangguo Sun, and Hong Cheng. 2025. When Do LLMs Help With Node Classification? A Comprehensive Analysis. In *Forty-second International Conference on Machine Learning*.
[56] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems* 32, 1 (2020), 4–24.
[57] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How Powerful are Graph Neural Networks?. In *International Conference on Learning Representations*.
[58] Hao Yan, Chaozhuo Li, Ruosong Long, Chao Yan, Jianan Zhao, Wenwen Zhuang, Jun Yin, Peiyan Zhang, Weihao Han, Hao Sun, et al. 2023. A comprehensive study on text-attributed graphs: Benchmarking and rethinking. *Advances in Neural Information Processing Systems* 36 (2023), 17238–17264.
[59] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. *Advances in neural information processing systems* 33 (2020), 5812–5823.
[60] Jianxiang Yu, Yuxiang Ren, Chenghua Gong, Jiaqi Tan, Xiang Li, and Xuecang Zhang. 2025. Leveraging large language models for node generation in few-shot learning on text-attributed graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 13087–13095.
[61] Mengmei Zhang, Mingwei Sun, Peng Wang, Shen Fan, Yanhu Mo, Xiaoxiao Xu, Hong Liu, Cheng Yang, and Chuan Shi. 2024. Graphtranslator: Aligning graph model to large language model for open-ended tasks. In *Proceedings of the ACM Web Conference 2024*. 1003–1014.
[62] Zhongjian Zhang, Xiao Wang, Huichi Zhou, Yue Yu, Mengmei Zhang, Cheng Yang, and Chuan Shi. 2025. Can large language models improve the adversarial robustness of graph neural networks?. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*. 2008–2019.
[63] Huachi Zhou, Jiahe Du, Chuang Zhou, Chang Yang, Yilin Xiao, Yuxuan Xie, and Xiao Huang. 2025. Each Graph is a New Language: Graph Learning with LLMs. *arXiv preprint arXiv:2501.11478* (2025).

A Datasets

We summarize the basic statistics of the datasets in Table 7. **Citation Networks**: **Arxiv** [20] is a large-scale CS paper citation network with 40 subject categories; **Pubmed** [18] is a biomedical citation network with 3 disease-related classes; **Cora** [52] is a machine learning citation graph with 70 fine-grained categories on the provided subset. **E-commerce Graphs**: **Computer** [58] is a co-purchase graph of computer products with 10 categories; **Books-children** and **Books-History** [58] are co-purchase graphs of children's and history books with 24 and 12 categories, respectively; **Photo** and **Sports** [58] are co-purchase graphs of photography and fitness products with 12 and 13 categories, respectively. **Web Link Graph**: **WikiCS** [36] is a Wikipedia CS article graph with 10 sub-fields. **Social Networks**: **Instagram** and **Reddit** [24] are user interaction graphs with 2 classes each. For all datasets we follow the split setting in TEA-GLM [47].

B Prompt Design

Each graph-language task in MOBI is formulated as a unified textual instruction comprising four components: (1) **Graph Context** (`<graph>`) — the subgraph centered on the target node; (2) **Textual Attributes** (`{{raw_text}}`) — domain-specific text such as titles, abstracts, or product reviews; (3) **Task Question** — a natural language query specifying the prediction objective; and (4) **Answer Candidates** (`{{label_names}}`) — a closed-form label set for constrained generation. A representative template is shown below.

Table 7: Statistics of the datasets.

Domain	Dataset	Nodes	Edges	Classes
Citation	Arxiv [20]	169,343	1,166,243	40
	Pubmed [18]	19,717	44,338	3
	Cora [52]	25,120	91,140	70
E-commerce	Computer [58]	87,229	721,081	10
	Books-children [58]	76,875	1,554,578	24
	Books-History [58]	41,551	358,574	12
	Photo [58]	48,362	500,928	12
	Sports [58]	173,055	1,773,500	13
Web link	WikiCS [36]	11,701	216,123	10
Social network	Instagram [24]	11,339	144,010	2
	Reddit [24]	33,434	198,438	2

Table 8: Key hyperparameters of MOBI.

Hyperparameter	Value Range / Setting
Number of dual-pathway layers (K)	{2, 4, 6, 8, 10}
Bias magnitude (b)	{0, 1, 2, 3, 4}
Perturbation ratio (p)	{15%, 40%, 80%}
Training epochs	2
Learning rate	5×10^{-5}
Batch size	16
Optimizer	AdamW

Given a citation graph from arXiv Computer Science papers: `<graph>` where the first node is the target paper, and other nodes are its one-hop or multi-hop neighbors, with the following information: `{{raw_text}}` Question: Which arXiv CS sub-category does this target paper belong to? Please directly give the most likely answer from the following subcategories: `{{label_names}}`.

C Experimental Details

Software Configuration. We implement our method using PyTorch 2.8.0, PyG 2.5.0 and Transformers 4.57.2. For GraphGPT and GOFA, we re-run their official implementations under our setting. For ZeroG, we utilize the implementations provided in [55]. We thank the authors for their reliable implementations.

Hyperparameter. Table 8 summarizes the key hyperparameters used in MOBI training and evaluation.

D Overall Algorithm

Algorithm 1 presents the overall training and inference procedure of MOBI, consolidating the three key components described in Sections 4.

E Additional Experimental Results

E.1 Additional Results on Hallucination Rates

To further examine the factual consistency and structural grounding of our approach, we compare the hallucination rates between the baseline and our proposed model. As illustrated in Table 9, the proposed model achieves pronounced reductions in hallucination frequency on four out of six evaluation sets. These results indicate that MOBI's unified graph–language modeling coupled with dual-pathway parameterization and progressive interaction scheduling

Algorithm 1: Overall Process of MOBI

Input: Pretraining graph $\mathcal{D}_{pre}$;
Parameter: LLM with L layers; number of dual-pathway blocks K; modality-wise bias function $b_c(\cdot)$; perturbation ratio p;
Output: Predicted labels for downstream tasks;
// Preprocessing
Encode node attributes via sentence encoder;
Precompute graph LPE encodings and shortest-path distances;
// Training on $\mathcal{D}_{pre}$
for *each training step* **do**
  Apply correlation-guided attribute perturbation;
  // Forward pass
  for $\ell = 1, \cdots, K$ (dual-pathway blocks) **do**
    Route tokens to separate pathways via Eq. (2-3), in which attention score is computed via Eq. (1) and Eq. (4);
  for $\ell = K{+}1, \cdots, L$ (normal blocks) **do**
    Apply standard transformer layer, in which attention score is computed via Eq. (1);
  Compute loss; update parameters;
// Predicting on $\mathcal{D}_{down}$
Forward pass through MOBI (no perturbation);
return Decoded predictions;

Table 9: Comparison of hallucination rates between the baseline model and the proposed model.

Model	PubMed	Cora-large	Children	History	Photo	Sports
Vicuna-7B-v1.5	0.0013	0.1061	0.0148	0.1332	0.0258	0.0343
MOBI	0.0003	0.0967	0.0368	0.0053	0.0110	0.0254

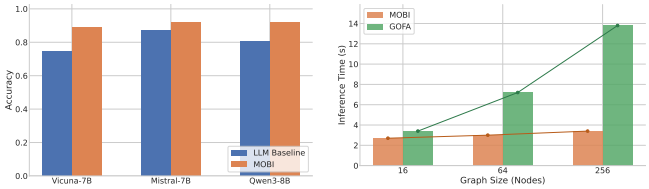
effectively curtails the generation of unsupported or semantically inconsistent content. This capability is especially critical for zero-shot cross-domain scenarios.

E.2 Performance with various LLM Backbones

Vicuna-7B was chosen as the default LLM for fair baseline comparisons. To verify the generalizability of our method, we evaluate MOBI with two additional LLMs (Mistral-7B and Qwen3-8B). Figure 6a presents the results on Pubmed. Notably, while stronger LLMs (Mistral, Qwen3) achieve higher baseline scores, MOBI maintains a significant performance margin in all cases. This demonstrates that MOBI's core advantages, i.e., superior structural understanding and mitigated modality interference, are orthogonal to the capacity of the underlying LLM backbone.

E.3 Additional Efficiency Analysis

Parameter Efficiency. A natural concern with MOBI's dual-pathway design is whether it introduces significant parameter overhead. We clarify that the graph pathway is restricted to only the first K layers



(a) Performance with different LLM backbones on Pubmed.

(b) Inference time (s) with different graph size.

(default $K = 6$) of the Transformer backbone. Compared to the entire LLM backbone, this decoupling introduces only a small number of additional parameters. Meanwhile, recent sota GFMs, such as GOFA and UniGTE, employ an entirely separate transformer as an external graph encoder. In contrast, our monolithic design completely eliminates the need for such heavyweight external modules, utilizing fewer overall parameters (MOBI's 7.91B vs. GOFA's 9.62B).

Inference Efficiency. Figure 6b reports the inference time (seconds) under varying input graph size. MOBI directly inputs node embeddings as graph tokens, making inference time increase marginal. Conversely, GOFA scales poorly as it encodes raw text per node.

E.4 Additional Results on Graph-level Tasks

To further demonstrate the versatility of MOBI, we extend our evaluation to graph-level tasks. By simply adapting the instruction template, MOBI can be seamlessly applied to graph classification. We conduct experiments on three widely used molecular graph classification benchmarks: BBBP, BACE, and HIV, under a supervised setting. We compare MOBI against the fine-tuned pure LLM baseline, which takes SMILES strings as input without access to explicit graph structure.

As shown in Table 10, MOBI consistently outperforms the fine-tuned pure LLM baseline across all three benchmarks. This demonstrates that MOBI successfully captures graph-level topological information that pure LLMs struggle to learn from SMILES strings alone. We leave large-scale zero-shot evaluation on graph-level tasks and Graph Question Answering (GraphQA) for future work.

Table 10: Graph classification results (ROC-AUC).

Model	BBBP	BACE	HIV
Vicuna-7B-v1.5	0.577	0.561	0.641
MOBI	**0.632**	**0.638**	**0.703**