UNICO: FOSTERING CROSS-DOMAIN AND CROSS-MODAL GENERALIZATION FOR GRAPH FOUNDATION MODELS VIA UNIFIED COVARIANCE OPERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Input feature heterogeneity across domains severely impairs the transferability of graph foundation models. Although natural language instructions provide task priors, existing cross-domain heterogeneous graph models are restricted to the graph-only modality, while current graph-text models suffer from either topological distortion induced by graph serialization or rigid dependency on domain-specific alignment. To this end, we propose UniCO, a cross-domain graph-text framework based on unified covariance operators. UniCO conceptualizes text instructions as "text graphs", thereby casting the dual cross-domain and cross-modal challenge into a singular heterogeneous graph generalization problem. Building upon this, we construct unified multi-order covariance operators for fine-grained cross-modal interactions, and design an asymmetric mask to preserve instruction temporal order while isolating the graph from instruction noise. Theoretically, we prove that these operators exhibit joint distributional and expectation invariance under input feature basis transformations, and reconcile "macroscopic modal decoupling" with "microscopic graph-text interactions" in expectation, eliminating dependence on domain-specific feature spaces and laying a theoretical foundation for cross-domain transfer. Experiments on multiple molecular graph and text-attributed graph benchmarks demonstrate that UniCO achieves superior zero-shot cross-domain generalization and efficient few-shot fine-tuning performance. Our code is available at https://anonymous.4open.science/r/UniCO-1923/.

1 INTRODUCTION

Existing graph foundation models (GFMs) are predominantly pre-trained and fine-tuned on single-domain graphs, struggling to generalize and reason across diverse domains Mao et al. (2024). Unlike computer vision (which shares standardized RGB formats) or natural language processing (which operates on unified token vocabularies), real-world graphs lack a common input feature space. Consequently, graphs from disparate domains are heterogeneous and incompatible in feature dimensionality, physical scale, numerical distribution, and basis semantics Frasca et al. (2024); Eliasof et al. (2026), fundamentally limiting the cross-domain transferability of conventional GFMs.

Recent efforts have introduced natural language instructions to guide the zero-shot and few-shot transfer of GFMs. However, existing graph-text models suffer from architectural and representational deficiencies. Some studies Chen et al. (2024a); Shirai et al. (2025); Gong et al. (2026) flatten graphs into token sequences to accommodate language models (LMs), which irreversibly destroys higher-order connectivity. Others Liu et al. (2024); Tang et al. (2024); He et al. (2025); Xu et al. (2025); Wu et al. (2026) employ decoupled encoders for graphs and text to preserve modality-specific features, but the global dense vectors compressed by graph encoders restrict the LM's perception of local structural details. Still others Zhao et al. (2023); Jing et al. (2026) support unified graph-text encoding yet heavily rely on complex cross-modal attention or domain-specific alignment layers, which inevitably collapse once transferred to target domains with unseen feature spaces.

To enhance the generalization of GFMs on cross-domain heterogeneous graphs, recent studies have attempted to align feature spaces via techniques such as specialized encoders Lachi et al. (2024); Sun et al. (2025); Yu et al. (2025), structuralization Xia et al. (2024); Wang et al. (2025); Sun et al. (2026),

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.

feature projections Zhao et al. (2024); Yuan et al. (2025), and order statistics Shen et al. (2025); Chen et al. (2026). Nevertheless, these approaches typically require careful adaptation to new scenarios, precluding end-to-end generalization. To mitigate this, recent works have constructed cross-domain graph models based on second-order statistics Eliasof et al. (2026). Yet, they are restricted to the pure-graph modality, failing to enable instruction-oriented multimodal representation learning.

In this work, we propose a reverse modeling paradigm. Instead of serializing graphs into text, we conceptualize unstructured instructions as "text graphs". Specifically, we treat each instruction token as a node and construct a "text adjacency matrix" governed by the information flow. Consequently, we obtain a unified graph composed of real-world heterogeneous graphs and "text graphs" across diverse domains, thereby reducing the intricate cross-domain and cross-modal dual challenge into a singular cross-domain heterogeneous graph generalization problem. We construct parameter-free unified multi-order covariance operators that enable fine-grained cross-domain and cross-modal interactions while theoretically eliminating dependence on domain-specific feature spaces.

Our contributions can be summarized as follows: (1) We propose an innovative "text graph" modeling paradigm that abstracts intricate cross-domain and cross-modal interactions into a singular heterogeneous graph generalization problem; (2) We construct multi-order covariance operators and prove that they enforce "modal decoupling" in expectation while capturing "graph-text co-variation" at the instance level, providing a theoretical foundation for cross-domain and cross-modal transfer; (3) We design an asymmetric mask to regularize the covariance operators, allowing instructions to perceive both temporal context and graph structure while shielding the graph from instruction noise.

2 Related Work

Graph-Text Models. Existing works mainly fall into three streams. First, pure-text reasoning models, such as NAG Gong et al. (2026) and GraphToken Perozzi et al., linearize graphs into text sequences for LLM reasoning, inevitably sacrificing native higher-order connectivity. Second, dual-encoder frameworks align multimodal features via contrastive learning (e.g., GraphCLIP Zhu et al. (2025)) or construct prompt graphs and virtual node connections (e.g., GraphGPT Tang et al. (2024), UniGraph He et al. (2025), OFA Liu et al. (2024)). However, the global embeddings output by unimodal encoders hinder fine-grained cross-modal perception. Third, interactive graph-text models, such as GIMLET Zhao et al. (2023), LLaGA Chen et al. (2024b), and EDT-Former Jing et al. (2026), rely on domain-specific attention interaction layers, precluding direct cross-domain generalization.

Cross-Domain Graph Models. Existing efforts pursue three directions. First, feature-mapping models align feature distributions across domains. GCOPE Zhao et al. (2024) introduces coordination nodes to bridge heterogeneous topologies, AnyGraph Xia & Huang (2026) and BRIDGE Yuan et al. (2025) employ mixture-of-experts networks for knowledge transfer, while AutoGFM Chen et al. (2025) tailors architectures to domain inductive biases. Second, topology-regularization models tackle topological heterogeneity. SAMGPT Yu et al. (2025) uses structural prompts to disentangle shared and specific features, while RiemannGFM Sun et al. (2025) and GraphGlue Sun et al. (2026) map topologies onto non-Euclidean spaces to capture structural invariants via manifold stitching and geometric operators. Yet, they typically demand re-adaptation to unseen scenarios. Third, statistics-based methods, such as ALLIN Eliasof et al. (2026), design invariant statistical operators to eliminate feature-basis dependencies, yet fail to exploit task priors in instructions.

3 Method: UniCO

This section presents the construction pipeline of UniCO. We formalize the predictive paradigm in Sec. 3.1, elaborate on the model architecture in Sec. 3.2, and specify the pre-training optimization objective in Sec. 3.3. The overall architecture of UniCO is illustrated in Figure 1.

3.1 Problem Formulation

We formulate diverse graph prediction tasks into an instruction-guided generative paradigm over entire graphs or induced subgraphs. For a specific graph task $\tau \in \mathcal{T}$, each instance is formulated as a triplet $(\mathcal{G}, \boldsymbol{T}^\tau, \boldsymbol{y}^\tau)$. Formally, let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A}_g)$ denote a graph, where $\mathcal{V} = \{v_1, \ldots, v_{|\mathcal{V}|}\}$

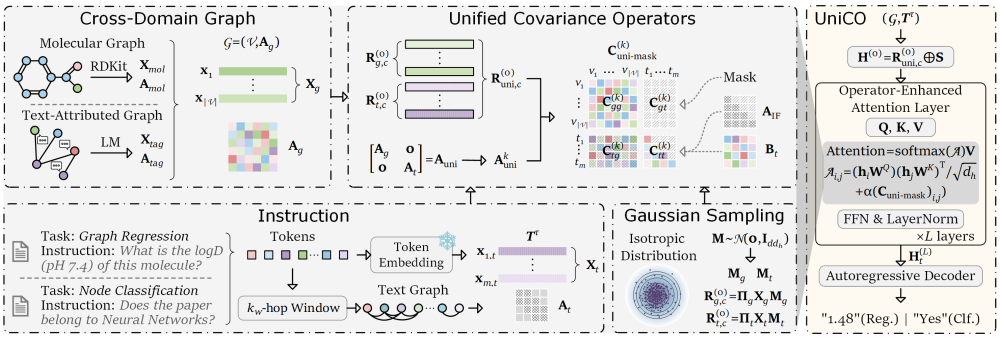


Figure 1: Overview of the UniCO construction pipeline.

is the node set, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the set of directed or undirected edges, and $\mathbf{A}_g \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ is the adjacency matrix. The node feature matrix is denoted by $\mathbf{X}_g = [\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathcal{V}|}]^\top \in \mathbb{R}^{|\mathcal{V}|\times d_n}$, where $\mathbf{x}_i \in \mathbb{R}^{d_n}$ denotes the feature vector of v_i. If available, the edge feature matrix is given by $\mathbf{X}_e = [\mathbf{e}_1, \ldots, \mathbf{e}_{|\mathcal{E}|}]^\top \in \mathbb{R}^{|\mathcal{E}|\times d_e}$, with $\mathbf{e}_i \in \mathbb{R}^{d_e}$ representing the attribute vector of the i-th edge. The instruction $\boldsymbol{T}^\tau$ injects prior knowledge and task queries into the model. It is tokenized into a sequence $(t_1, \ldots, t_m)$ and embedded into a token matrix $\mathbf{X}_t = [\mathbf{x}_{1,t}, \ldots, \mathbf{x}_{m,t}]^\top \in \mathbb{R}^{m\times d_t}$ via the LM's embedding lookup table, where $\mathbf{x}_{j,t} \in \mathbb{R}^{d_t}$ is the semantic vector of the j-th token. We map the prediction targets of arbitrary graph tasks to a text sequence $\boldsymbol{y}^\tau = (u_1, \ldots, u_S)$ of length $|\boldsymbol{y}^\tau| = S$, with each token $u_s \in \mathcal{V}_{\text{vocab}}$. Taking the joint graph constructed from $\mathcal{G}$ and $\boldsymbol{T}^\tau$ as input, the encoder extracts contextual representations to condition the autoregressive decoder to generate the target sequence $\hat{\boldsymbol{y}}^\tau = \text{UniCO}(\mathcal{G}, \boldsymbol{T}^\tau)$. The notations and definitions are listed in Appendix A.

3.2 Unified Graph-Text Covariance Architecture

Text Graph Construction. We adopt an encoder-decoder LM Colin (2020) as the backbone to construct the unified graph-text network. The model treats physical graph nodes and instruction tokens as individual "nodes", composing them into a joint graph. The first $|\mathcal{V}|$ nodes represent the graph $\mathcal{G}$ with node features $\mathbf{X}_g \in \mathbb{R}^{|\mathcal{V}|\times d_n}$ and adjacency matrix $\mathbf{A}_g \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$, while the subsequent m nodes denote the instruction $\boldsymbol{T}^\tau$ (forming a "text graph") with token matrix $\mathbf{X}_t \in \mathbb{R}^{m\times d_t}$ and adjacency matrix $\mathbf{A}_t \in \mathbb{R}^{m\times m}$. Specifically, $\mathbf{A}_t$ is built via a sliding window of size k_w, where $(\mathbf{A}_t)_{i,j} = 1$ if the token pair (t_i, t_j) satisfies $|i-j| \le k_w$, and 0 otherwise, thus retaining local temporal dependencies in $\boldsymbol{T}^\tau$. Both $\mathbf{A}_g$ and $\mathbf{A}_t$ are symmetrically normalized. Consequently, the unified adjacency matrix is initialized as $\mathbf{A}_{\text{uni}} = \text{diag}(\mathbf{A}_g, \mathbf{A}_t) \in \mathbb{R}^{N\times N}$, where $N = |\mathcal{V}| + m$.

Stochastic Feature Projections. At each forward pass, the model stochastically samples two isotropic Gaussian random matrices $\mathbf{M}_g \in \mathbb{R}^{d_n\times d_h}$ and $\mathbf{M}_t \in \mathbb{R}^{d_t\times d_h}$ with $\text{vec}(\mathbf{M}_g) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_n d_h})$ and $\text{vec}(\mathbf{M}_t) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_t d_h})$. They project heterogeneous graphs from disparate domains and "text graphs" from the semantic space into a shared latent feature space $\mathbb{R}^{d_h}$:

$$\mathbf{R}_g^{(0)} = \mathbf{X}_g\mathbf{M}_g \in \mathbb{R}^{|\mathcal{V}|\times d_h}, \quad \mathbf{R}_t^{(0)} = \mathbf{X}_t\mathbf{M}_t \in \mathbb{R}^{m\times d_h}. \tag{1}$$

As proven in Proposition 4.1, the projected features $\mathbf{R}_g^{(0)}$ and $\mathbf{R}_t^{(0)}$ are invariant in distribution to feature-axis transformations of their respective input spaces, thereby eliminating dependence on domain-specific and modality-specific coordinate bases.

Unified Covariance Operators. We apply the geometric centering operators $\mathbf{\Pi}_g = \mathbf{I}_{|\mathcal{V}|} - \frac{1}{|\mathcal{V}|}\mathbf{1}_{|\mathcal{V}|}\mathbf{1}_{|\mathcal{V}|}^\top \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ and $\mathbf{\Pi}_t = \mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top \in \mathbb{R}^{m\times m}$ (where $\mathbf{1}$ denotes the all-ones column vector) to $\mathbf{R}_g^{(0)}$ and $\mathbf{R}_t^{(0)}$, respectively, and concatenate the centered features $\mathbf{R}_{g,c}^{(0)}$ and $\mathbf{R}_{t,c}^{(0)}$ to formulate the unified features $\mathbf{R}_{\text{uni},c}^{(0)}$:

$$\mathbf{R}_{\text{uni},c}^{(0)} = \begin{bmatrix}\mathbf{R}_{g,c}^{(0)} \\ \mathbf{R}_{t,c}^{(0)}\end{bmatrix} = \begin{bmatrix}\mathbf{\Pi}_g\mathbf{R}_g^{(0)} \\ \mathbf{\Pi}_t\mathbf{R}_t^{(0)}\end{bmatrix} = \begin{bmatrix}\mathbf{\Pi}_g\mathbf{X}_g\mathbf{M}_g \\ \mathbf{\Pi}_t\mathbf{X}_t\mathbf{M}_t\end{bmatrix} \in \mathbb{R}^{N\times d_h}. \tag{2}$$

Based on $\mathbf{R}^{(0)}_{\text{uni},c}$, we compute the unified zero-order node covariance operator $\mathbf{C}^{(0)}_{\text{uni}}$:

$$\mathbf{C}^{(0)}_{\text{uni}} = \frac{1}{d_h}\mathbf{R}^{(0)}_{\text{uni},c}(\mathbf{R}^{(0)}_{\text{uni},c})^\top = \frac{1}{d_h}\begin{bmatrix}\mathbf{R}^{(0)}_{g,c}(\mathbf{R}^{(0)}_{g,c})^\top & \mathbf{R}^{(0)}_{g,c}(\mathbf{R}^{(0)}_{t,c})^\top \\ \mathbf{R}^{(0)}_{t,c}(\mathbf{R}^{(0)}_{g,c})^\top & \mathbf{R}^{(0)}_{t,c}(\mathbf{R}^{(0)}_{t,c})^\top\end{bmatrix} = \begin{bmatrix}\mathbf{C}^{(0)}_{gg} & \mathbf{C}^{(0)}_{gt} \\ \mathbf{C}^{(0)}_{tg} & \mathbf{C}^{(0)}_{tt}\end{bmatrix}, \tag{3}$$

where the off-diagonal blocks $\mathbf{C}^{(0)}_{gt} \in \mathbb{R}^{|\mathcal{V}|\times m}$ and $\mathbf{C}^{(0)}_{tg} \in \mathbb{R}^{m\times|\mathcal{V}|}$ naturally serve as parameter-free, statistically grounded cross-modal conduits. Bypassing dedicated alignment layers, they quantify the statistical co-variation between graphs and instructions in the latent space. Meanwhile, the diagonal blocks $\mathbf{C}^{(0)}_{gg} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ and $\mathbf{C}^{(0)}_{tt} \in \mathbb{R}^{m\times m}$ capture intrinsic pairwise geometric similarities within the graph and text modalities, respectively, effectively acting as dense relational graphs.

To incorporate multi-scale structural and contextual dependencies, we propagate $\mathbf{R}^{(0)}_{\text{uni},c}$ over $\mathbf{A}_{\text{uni}}$. For any propagation order $k \in \{0,\dots,K\}$, the centered features are formalized as $\mathbf{R}^{(k)}_{\text{uni},c} = [(\mathbf{\Pi}_g\mathbf{A}^k_g\mathbf{R}^{(0)}_{g,c})^\top,\ (\mathbf{\Pi}_t\mathbf{A}^k_t\mathbf{R}^{(0)}_{t,c})^\top]^\top$, yielding the k-order unified covariance operator $\mathbf{C}^{(k)}_{\text{uni}} = \frac{1}{d_h}\mathbf{R}^{(k)}_{\text{uni},c}(\mathbf{R}^{(k)}_{\text{uni},c})^\top$. Although graph nodes and instruction tokens are topologically isolated within $\mathbf{A}_{\text{uni}}$, the k-step propagation enables the cross-modal covariance operators $\mathbf{C}^{(k)}_{tg} = \frac{1}{d_h}\mathbf{R}^{(k)}_{t,c}(\mathbf{R}^{(k)}_{g,c})^\top = \frac{1}{d_h}(\mathbf{\Pi}_t\mathbf{A}^k_t\mathbf{R}^{(0)}_{t,c})(\mathbf{\Pi}_g\mathbf{A}^k_g\mathbf{R}^{(0)}_{g,c})^\top$ and $\mathbf{C}^{(k)}_{gt} = \frac{1}{d_h}\mathbf{R}^{(k)}_{g,c}(\mathbf{R}^{(k)}_{t,c})^\top = \frac{1}{d_h}(\mathbf{\Pi}_g\mathbf{A}^k_g\mathbf{R}^{(0)}_{g,c})(\mathbf{\Pi}_t\mathbf{A}^k_t\mathbf{R}^{(0)}_{t,c})^\top$ to automatically establish high-order statistical correlations between k-hop graph neighborhoods and k-step instruction contexts.

Asymmetric Masking Mechanism. Because $\mathbf{C}^{(k)}_{\text{uni}}$ is mathematically symmetric, directly deploying it as an unconstrained structural bias introduces two principal issues: (1) the dense connectivity of $\mathbf{C}^{(k)}_{tt}$ disrupts the information flow of the instruction, distorting its sequential dependencies; (2) the backward pathway $\mathbf{C}^{(k)}_{gt}$ permits the graph to attend to the instruction, potentially injecting task-specific instruction noise into graph representations during cross-domain generalization.

To resolve these issues, we devise an asymmetric masking mechanism. For $\mathbf{C}^{(k)}_{tt}$, we introduce a lower-triangular information flow matrix $\mathbf{A}_{\text{IF}} \in \mathbb{R}^{m\times m}$ and the LM's native relative positional bias matrix $\mathbf{B}_t \in \mathbb{R}^{m\times m}$, ensuring tokens attend only to preceding contexts while retaining sequential distance penalties. For $\mathbf{C}^{(k)}_{gt}$, we enforce a unidirectional blocking mask that shields the graph from perceiving the instruction. The modulated unified operator is formulated as:

$$\mathbf{C}^{(k)}_{\text{uni-mask}} = \begin{bmatrix}\mathbf{C}^{(k)}_{gg} & -\infty_{|\mathcal{V}|\times m} \\ \mathbf{C}^{(k)}_{tg} & \mathbf{C}^{(k)}_{tt} \oplus \mathbf{A}_{\text{IF}} \oplus \mathbf{B}_t\end{bmatrix}, \tag{4}$$

where $\oplus$ denote the element-wise addition. Finally, we aggregate the self-loop matrix $\mathbf{I}_{\text{uni}} = \text{diag}(\mathbf{I}_g, \mathbf{I}_t)$, the adjacency matrix $\mathbf{A}_{\text{uni}}$, and the modulated multi-order operators to yield the unified covariance operator set $\mathcal{C}_{\text{uni}} = \{\mathbf{I}_{\text{uni}}, \mathbf{A}_{\text{uni}}, \mathbf{C}^{(0)}_{\text{uni-mask}}, \dots, \mathbf{C}^{(K)}_{\text{uni-mask}}\}$.

Edge Covariance Operator. To accommodate graph domains with heterogeneous edge attributes $\mathbf{X}_e \in \mathbb{R}^{|\mathcal{E}|\times d_e}$, we project $\mathbf{X}_e$ via a Gaussian random matrix $\mathbf{M}_e \in \mathbb{R}^{d_e\times d_h}$ and aggregate the projected features of all incoming edges incident to each node, thereby deriving the edge-induced initial node features $\mathbf{R}^{(0)}_e \in \mathbb{R}^{|\mathcal{V}|\times d_h}$. Based on $\mathbf{R}^{(0)}_e$, we formulate the k-order edge-covariance operator $\mathbf{C}^{(k)}_e = \frac{1}{d_h}(\mathbf{\Pi}_g\mathbf{A}^k_g\mathbf{\Pi}_g\mathbf{R}^{(0)}_e)(\mathbf{\Pi}_g\mathbf{A}^k_g\mathbf{\Pi}_g\mathbf{R}^{(0)}_e)^\top \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$. Since "text graphs" lack native edge attributes, we extend this operator to the unified domain via zero-padding $\mathbf{C}^{(k)}_{\text{uni},e} = \text{diag}(\mathbf{C}^{(k)}_e, \mathbf{0}_{m\times m}) \in \mathbb{R}^{N\times N}$, which quantifies the statistical co-variation of edge environments between node pairs. This forms the operator set $\mathcal{C}_{\text{uni},e} = \{\mathbf{0}_{\text{uni}}, \mathbf{0}_{\text{uni}}, \mathbf{C}^{(0)}_{\text{uni},e}, \dots, \mathbf{C}^{(K)}_{\text{uni},e}\}$.

Operator-Enhanced Attention. We incorporate the unified operator sets $\mathcal{C}_{\text{uni}}$ and $\mathcal{C}_{\text{uni},e}$ as structural attention biases across independent attention heads. For the k-th head ($k \in \{0,\dots,K\}$) at layer l, the attention score between unified nodes i and j is formulated as:

$$\mathcal{A}^{(l,k)}_{i,j} = \frac{(\mathbf{h}^{(l-1)}_i\mathbf{W}^Q_k)(\mathbf{h}^{(l-1)}_j\mathbf{W}^K_k)^\top}{\sqrt{d_h}} + \alpha_k(\mathcal{C}_{\text{uni}}[k])_{i,j} + \beta_k(\mathcal{C}_{\text{uni},e}[k])_{i,j}, \tag{5}$$

where $\mathbf{W}^Q_k, \mathbf{W}^K_k$ are the query and key projection matrices for the k-th head, $\mathcal{C}_{\text{uni}}[k]$ and $\mathcal{C}_{\text{uni},e}[k]$ denote the corresponding operators indexed by the head, and α_k, β_k are learnable scaling coefficients.

The initial hidden features are formed as $\mathbf{H}^{(0)} = \mathbf{R}^{(0)}_{\text{uni},c} \oplus \mathbf{S}$, where $\mathbf{S} = \text{diag}(\mathbf{S}_g, \mathbf{S}_t) \in \mathbb{R}^{N \times d_h}$ integrates structural and positional encodings. $\mathbf{S}_g$ provides graph structural encodings (e.g., node degrees or random walk structural encodings), whereas $\mathbf{S}_t$ represents token positional encodings. Notably, $\mathbf{S}_t$ is not explicitly added to $\mathbf{H}^{(0)}$, but is implicitly injected via $\mathbf{B}_t$ within $\mathbf{C}^{(k)}_{\text{uni-mask}}$.

Autoregressive Text Decoding. The L-th (final) encoder layer outputs the unified hidden features $\mathbf{H}^{(L)} \in \mathbb{R}^{N \times d_h}$. Since the cross-modal block $\mathbf{C}_{tg}$ allows instruction tokens to thoroughly perceive structural information from the graph, we extract the features corresponding to text tokens, $\mathbf{H}^{(L)}_t = \mathbf{H}^{(L)}[|\mathcal{V}| + 1 : N, :] \in \mathbb{R}^{m \times d_h}$, which are then fed into the autoregressive decoder as cross-attention memory (keys and values) to generate the target text sequence.

3.3 Pre-Training Optimization Objective

The model is jointly pre-trained on the cross-domain task set $\mathcal{T} = \{\tau\}$, where each task τ is formulated as a conditional sequence generation objective. Given the ground-truth target text sequence $\boldsymbol{y}^\tau_i = (u^\tau_{i,1}, u^\tau_{i,2}, \ldots, u^\tau_{i,S^\tau_i})$ of length $S^\tau_i = |\boldsymbol{y}^\tau_i|$, variable target lengths across tasks can cause longer sequences to dominate gradient updates. Therefore, we adopt a sequence-length-normalized conditional autoregressive loss:

$$\mathcal{L}_{\text{pretrain}}(\Theta) = \frac{1}{\sum_{\tau \in \mathcal{T}} |\tau|} \sum_{\tau \in \mathcal{T}} \sum_{(\mathcal{G}_i, \boldsymbol{T}^\tau_i, \boldsymbol{y}^\tau_i) \in \tau} [-\frac{1}{S^\tau_i} \sum_{s=1}^{S^\tau_i} \log \mathbb{P}(u = u^\tau_{i,s} \mid u^\tau_{i,<s}, \mathbf{H}^{(L)}_t(\mathcal{G}_i, \boldsymbol{T}^\tau_i))], \quad (6)$$

where $|\tau|$ denotes the sample size of task τ, and Θ denotes the trainable parameters of the unified encoder and autoregressive decoder.

4 Theoretical Analysis

This section establishes the mathematical foundations of UniCO, providing rigorous theoretical justification for its cross-domain generalization (Sec. 4.1, Sec. 4.2), cross-modal interactions (Sec. 4.2), multi-order propagation stability (Sec. 4.3), and stochastic optimization convergence (Sec. 4.4).

4.1 Geometric Preservation and Distributional Invariance

Let $\mathbf{Q}_g \in \mathbb{R}^{d_n \times d_n}$ and $\mathbf{Q}_t \in \mathbb{R}^{d_t \times d_t}$ denote arbitrary orthogonal transformations acting on the graph node and instruction token feature spaces, respectively, satisfying $\mathbf{Q}_g \mathbf{Q}_g^\top = \mathbf{I}_{d_n}$ and $\mathbf{Q}_t \mathbf{Q}_t^\top = \mathbf{I}_{d_t}$. Let $\overline{\mathbf{X}}_g = \mathbf{X}_g \mathbf{Q}_g$ and $\overline{\mathbf{X}}_t = \mathbf{X}_t \mathbf{Q}_t$ be the transformed features.

Proposition 4.1.1 (Metric Preservation of Gaussian Projections)**.** *Define the stochastic linear mapping for a node or token feature vector* $\mathbf{x}$ *as* $f(\mathbf{x}) = \frac{1}{\sqrt{d_h}} \mathbf{x}\mathbf{M}$*, where* $\mathbf{M} \in \mathbb{R}^{d \times d_h} (d \in \{d_n, d_t\})$ *is an isotropic Gaussian matrix. For any* $\epsilon \in (0, 1/2)$ *and failure tolerance* $\delta \in (0, 1)$*, if the shared latent dimension satisfies* $d_h \geq \frac{8 \ln N^2 + 8 \ln(1/\delta)}{\epsilon^2}$*, then for any pair of feature vectors* $\mathbf{x}_u, \mathbf{x}_v \in \mathbb{R}^d$*, f preserves pairwise Euclidean distances and inner products with probability at least* $1 - \delta$*:*

$$(1 - \epsilon)\|\mathbf{x}_u - \mathbf{x}_v\|_2^2 \leq \|f(\mathbf{x}_u) - f(\mathbf{x}_v)\|_2^2 \leq (1 + \epsilon)\|\mathbf{x}_u - \mathbf{x}_v\|_2^2, \quad (7)$$

$$|\langle f(\mathbf{x}_u), f(\mathbf{x}_v)\rangle - \langle \mathbf{x}_u, \mathbf{x}_v\rangle| \leq \mathcal{O}(\epsilon). \quad (8)$$

Proposition 4.1.2 (Joint Distributional Invariance to Orthogonal Transformations)**.** *Under arbitrary orthogonal basis transformations* $(\mathbf{Q}_g, \mathbf{Q}_t)$*, the unified features* $\overline{\mathbf{R}}^{(0)}_{uni,c}$ *and the covariance operator set* $\overline{\mathcal{C}}_{uni}$ *are equal in distribution to their original counterparts:*

$$\overline{\mathbf{R}}^{(0)}_{uni,c} \overset{d}{=} \mathbf{R}^{(0)}_{uni,c}, \quad \overline{\mathcal{C}}_{uni} \overset{d}{=} \mathcal{C}_{uni}. \quad (9)$$

These propositions demonstrate that UniCO achieves geometric fidelity in the shared latent space without distorting native topological structures, and the input signals, i.e., the projected features and covariance operators, are decoupled from domain-specific coordinate bases. This establishes a topologically faithful and distributionally robust foundation for cross-modal and cross-domain interactions. Detailed proofs are provided in Appendix B.1.

4.2 Expected Invariance and Macroscopic Modal Decoupling

Theorem 4.2.1 (Expected Invariance and Modality Decoupling)**.** *For any propagation order k, the unified covariance operator $\mathbf{C}_{uni}^{(k)}$ is invariant in expectation to arbitrary coordinate transformations $(\mathbf{Q}_g, \mathbf{Q}_t)$. The off-diagonal cross-modal blocks vanish in expectation, while the diagonal blocks recover the centered Gram matrices of the native features:*

$$\mathbb{E}_{\mathbf{M}_g,\mathbf{M}_t}[\mathbf{C}_{uni}^{(k)}] = \begin{bmatrix} \mathbf{\Pi}_g \mathbf{A}_g^k \mathbf{\Pi}_g \mathbf{X}_g \mathbf{X}_g^\top \Pi_g (\mathbf{A}_g^k)^\top \mathbf{\Pi}_g & \mathbf{0}_{|\mathcal{V}|\times m} \\ \mathbf{0}_{m\times|\mathcal{V}|} & \mathbf{\Pi}_t \mathbf{A}_t^k \mathbf{\Pi}_t \mathbf{X}_t \mathbf{X}_t^\top \mathbf{\Pi}_t (\mathbf{A}_t^k)^\top \mathbf{\Pi}_t \end{bmatrix}. \tag{10}$$

The operators $\mathbf{C}_{uni\text{-}mask}^{(k)}$ also remain invariant to coordinate transformations in expectation.

UniCO achieves statistical unification across disparate domains. First-order statistics (sample means) are algebraically eliminated by the centering projector $\mathbf{\Pi}$, while second-order statistics (covariances) converge in expectation to dimension-agnostic centered Gram matrices, eliminating discrepancies in feature distributions and dimensionalities across heterogeneous graphs. Furthermore, the cross-modal covariance operator reconciles macro-level modal decoupling with micro-level dynamic interactions. Macroscopically, $\mathbb{E}[\mathbf{C}_{gt}^{(k)}] = \mathbf{0}$ acts as an implicit regularizer that prevents modal contamination. Microscopically, individual stochastic realizations of $\mathbf{C}_{tg}^{(k)}$ at each forward pass are almost surely non-zero, directly capturing shared channel-wise co-variations to enable instance-level graph-text interactions. We provide a detailed proof and an in-depth discussion on the theoretical conditions for cross-domain transferability in Appendix B.2.

4.3 Cross-Modal Structural Perception and Spectral Norm Stability

Proposition 4.3.1 (Structural Expressiveness)**.** *For any propagation order k, the (i,j)-th scalar entry of the cross-modal operator $\mathbf{C}_{tg}^{(k)}$ measures the empirical cross-covariance between the k-step context of instruction token i and the k-hop neighborhood of graph node j in the shared latent space:*

$$(\mathbf{C}_{tg}^{(k)})_{i,j} = \frac{1}{d_h}(\sum_{p=1}^{m}(\mathbf{\Pi}_t \mathbf{A}_t^k \mathbf{\Pi}_t)_{i,p}(\mathbf{R}_{t,c}^{(0)})_p)(\sum_{q=1}^{|\mathcal{V}|}(\mathbf{\Pi}_g \mathbf{A}_g^k \mathbf{\Pi}_g)_{j,q}(\mathbf{R}_{g,c}^{(0)})_q)^\top. \tag{11}$$

Proposition 4.3.2 (Spectral Boundedness)**.** *The spectral norm (matrix 2-norm) of $\mathbf{C}_{tg}^{(k)}$ admits an absolute upper bound independent of the propagation order k:*

$$\|\mathbf{C}_{tg}^{(k)}\|_2 \le \frac{1}{d_h}\|\mathbf{R}_{t,c}^{(0)}\|_2 \|\mathbf{R}_{g,c}^{(0)}\|_2. \tag{12}$$

The operator $\mathbf{C}_{tg}^{(k)}$ captures statistical dependencies between instructions and graphs, enabling automated cross-modal interactions without learnable cross-attention or alignment layers. Furthermore, its spectral norm does not diverge with the order k, preventing feature overflow and numerical collapse during attention bias injection in the LM. Detailed proofs are provided in Appendix B.3.

4.4 Operator Consistency and Objective Upper Bound

Theorem 4.4.1 (Operator Convergence in Probability)**.** *As the shared latent dimension $d_h \to \infty$, the empirical operator $\mathbf{C}_{uni}^{(0)}$ converges in probability to its mathematical expectation:*

$$\mathbf{C}_{uni}^{(0)} \xrightarrow{\mathcal{P}} \mathbb{E}_{\mathbf{M}_g,\mathbf{M}_t}[\mathbf{C}_{uni}^{(0)}] = \begin{bmatrix} \mathbf{\Pi}_g \mathbf{X}_g \mathbf{X}_g^\top \mathbf{\Pi}_g & \mathbf{0}_{|\mathcal{V}|\times m} \\ \mathbf{0}_{m\times|\mathcal{V}|} & \mathbf{\Pi}_t \mathbf{X}_t \mathbf{X}_t^\top \mathbf{\Pi}_t \end{bmatrix}. \tag{13}$$

For any finite dimension d_h and failure tolerance $\delta \in (0,1)$, the relative spectral error satisfies with probability at least $1-\delta$:

$$\frac{\|\mathbf{C}_{gg}^{(0)} - \mathbf{\Pi}_g \mathbf{X}_g \mathbf{X}_g^\top \mathbf{\Pi}_g\|_2}{\|\mathbf{\Pi}_g \mathbf{X}_g \mathbf{X}_g^\top \mathbf{\Pi}_g\|_2} \le \mathcal{O}\left(\sqrt{\frac{r(\mathbf{\Pi}_g \mathbf{X}_g)}{d_h}} + \sqrt{\frac{\ln(1/\delta)}{d_h}}\right), \tag{14}$$

where $r(\mathbf{\Pi}_g \mathbf{X}_g) = \frac{\mathrm{Tr}(\mathbf{\Pi}_g \mathbf{X}_g \mathbf{X}_g^\top \mathbf{\Pi}_g)}{\|\mathbf{\Pi}_g \mathbf{X}_g \mathbf{X}_g^\top \mathbf{\Pi}_g\|_2}$ denotes the effective rank of $\mathbf{\Pi}_g \mathbf{X}_g$.

Table 1: Zero-shot performance (ROC-AUC ↑, mean % ± standard deviation %) of UniCO and baselines on molecular graph classification tasks. The best results are in **bold** and the second-best results are as underlined (This notation applies to all subsequent tables).

Task	Graph Classification								
Model	BACE	HIV	MUV	Tox21	ToxCast	BBBP	PCBA	ChEMBL-ZS	Avg. Graph-Clf
GraphMAE	52.30±0.46	50.82±0.21	53.13±0.28	51.53±0.53	50.20±0.37	51.69±0.42	49.85±0.22	46.54±0.31	50.76±2.01
OFA	61.56±0.34	55.67±0.26	58.22±0.48	54.30±0.29	52.11±0.41	56.00±0.33	60.62±0.27	53.28±0.44	56.47±3.40
MoMu	66.56±0.52	50.26±0.30	60.51±0.31	57.57±0.47	52.38±0.35	49.81±0.27	51.50±0.32	50.02±0.38	54.83±6.14
KVPLM	51.26±0.33	61.20±0.22	61.72±0.38	49.17±0.26	50.96±0.34	60.20±0.34	48.11±0.24	41.55±0.49	53.02±7.29
Galactica-1.3B	56.48±0.41	33.85±0.22	57.15±0.35	49.46±0.28	51.23±0.26	53.94±0.29	52.02±0.35	48.18±0.40	50.29±7.35
Llama-7B	56.83±0.42	48.14±0.29	57.31±0.24	55.76±0.33	56.15±0.34	54.21±0.38	54.03±0.37	57.51±0.38	54.99±3.06
Llama2-7B	58.55±0.47	53.57±0.34	58.18±0.36	57.22±0.35	57.85±0.29	54.84±0.30	55.67±0.21	58.19±0.25	56.76±1.84
GIMLET	69.57±0.26	66.24±0.30	64.39±0.17	61.19±0.43	59.04±0.26	59.39±0.31	62.11±0.25	78.60±0.47	65.07±6.52
UniGraph	66.89±0.33	67.24±0.34	64.50±0.28	58.94±0.48	56.25±0.33	58.81±0.27	52.40±0.36	55.83±0.24	60.11±5.50
UniCO	**80.86**±0.31	**73.24**±0.22	**66.03**±0.25	**63.02**±0.36	**60.15**±0.32	**60.52**±0.13	**63.68**±0.28	**80.43**±0.18	**68.49**±8.54

Table 3: Zero-shot performance (ROC-AUC ↑, mean % ± standard deviation %) of UniCO and baselines on text-attributed graph node classification and link prediction tasks.

Task	Node Classification					Link Prediction			
Model	Cora	CiteSeer	Instagram	WikiCS	Avg. Node-Clf	Cora_edge	CiteSeer_edge	WikiCS_edge	Avg. Link-Pred
GRACE	50.19±0.42	51.28±0.48	50.83±0.24	49.27±0.46	50.39±0.87	52.29±0.31	55.82±0.47	47.53±0.33	51.88±4.17
GraphMAE	52.84±0.35	50.15±0.36	**53.09**±0.48	51.43±0.30	51.88±1.36	53.17±0.34	56.53±0.40	48.89±0.25	52.86±3.87
ZeroG	60.12±0.43	56.22±0.39	50.57±0.21	56.53±0.37	55.86±3.95	55.80±0.28	60.48±0.34	51.14±0.39	55.81±4.67
OFA	56.85±0.29	53.51±0.34	49.61±0.37	54.52±0.45	53.62±0.34	58.56±0.31	58.25±0.42	50.97±0.28	55.93±4.29
G2P2	54.21±0.47	51.82±0.40	51.27±0.43	53.15±0.42	52.61±1.33	54.56±0.46	57.87±0.31	48.28±0.34	53.57±4.84
GraphCLIP	59.74±0.42	57.75±0.30	51.14±0.32	57.00±0.28	56.41±3.70	53.37±0.35	61.41±0.45	51.75±0.40	55.51±4.11
GraphGPT	51.94±0.43	51.45±0.39	51.42±0.43	50.10±0.37	51.23±0.79	52.80±0.42	56.28±0.49	50.66±0.38	53.25±2.82
Llama-7B	53.83±0.36	53.18±0.43	49.15±0.35	53.22±0.43	52.35±2.15	54.12±0.45	58.01±0.31	50.83±0.38	54.32±3.60
Llama2-7B	55.21±0.38	53.50±0.24	50.87±0.31	54.80±0.26	53.60±1.96	53.55±0.44	58.60±0.32	52.21±0.21	54.79±3.20
Llama3.1-8B	57.55±0.20	56.84±0.39	50.27±0.29	54.32±0.45	54.75±3.29	56.21±0.39	58.01±0.31	52.82±0.32	55.68±2.65
UniGraph	62.83±0.35	61.59±0.44	51.50±0.34	55.27±0.46	57.80±5.35	58.93±0.22	65.22±0.28	51.56±0.24	58.57±6.86
UniCO	**65.28**±0.27	**65.24**±0.36	50.78±0.42	**57.29**±0.38	**59.65**±7.00	**61.64**±0.24	**73.24**±0.30	**54.81**±0.29	**63.23**±9.34

Let $\mathbf{H}_t^{(L)} \in \mathbb{R}^{m \times d_h}$ denote the text token features output by the L-th encoder under a stochastic realization $(\mathbf{M}_g, \mathbf{M}_t)$. The autoregressive decoder models the conditional likelihood $\mathbb{P}(\hat{\boldsymbol{y}} = \boldsymbol{y}_{\text{true}} \mid \mathbf{H}_t^{(L)})$, evaluated via the loss $\mathcal{L}(\mathbb{P}(\hat{\boldsymbol{y}} = \boldsymbol{y}_{\text{true}} \mid \mathbf{H}_t^{(L)}), \boldsymbol{y}_{\text{true}})$, where $\boldsymbol{y}_{\text{true}}$ is the target text sequence.

Proposition 4.4.2 (Training Objective Upper Bound)**.** *Assuming that the objective function $\mathcal{L}(\mathbb{P}(\hat{\boldsymbol{y}} = \boldsymbol{y}_{true} \mid \mathbf{H}_t^{(L)}), \boldsymbol{y}_{true})$ is convex with respect to $\mathbf{H}_t^{(L)}$, the stochastic empirical training objective provides a rigorous upper bound on the loss of the expected stationary representation:*

$$\mathcal{L}(\mathbb{P}(\hat{\boldsymbol{y}} \mid \mathbb{E}_{\mathbf{M}_g,\mathbf{M}_t}[\mathbf{H}_t^{(L)}]), \boldsymbol{y}_{true}) \leq \mathbb{E}_{\mathbf{M}_g,\mathbf{M}_t}[\mathcal{L}(\mathbb{P}(\hat{\boldsymbol{y}} \mid \mathbf{H}_t^{(L)}), \boldsymbol{y}_{true})]. \tag{15}$$

In practice, moderate projection dimensions (e.g., $d_h = 512$, where $r(\mathbf{\Pi X}) \ll d_h$) are sufficient to guarantee that the empirical covariance operator approximates the centered Gram matrix with high probability, making zero-shot cross-domain transfer practically tractable. Crucially, optimizing the empirical objective over stochastic forward passes bounds the true predictive risk on the expected stationary representation, ensuring stable convergence during pre-training and cross-domain transfer. Detailed proofs are provided in Appendix B.4.

5 Experiment

5.1 Experimental Setup

To systematically evaluate the generalization of UniCO across diverse domains, modalities, and graph tasks, our experiments include 20 datasets spanning two benchmarks: molecular graphs and text-attributed graphs. The baselines comprise 15 representative methods, categorized into self-supervised graph models, topological prompt models, dual-encoder contrastive models, and generative multimodal LMs. For

Table 2: Zero-shot performance (RMSE ↓, mean ± standard deviation) of UniCO and baselines on molecular graph regression tasks.

Task	Graph Regression			
Model	ESOL	Lipo	FreeSolv	Avg. Graph-Reg
GraphMAE	2.46±0.02	1.82±0.01	6.85±0.02	3.71±2.74
Galactica-1.3B	2.16±0.02	1.78±0.02	6.20±0.03	3.38±2.45
Llama-7B	2.68±0.02	2.13±0.03	6.35±0.02	3.72±2.29
Llama2-7B	2.56±0.01	2.02±0.03	6.15±0.01	3.58±2.24
GIMLET	**1.13**±0.02	1.35±0.02	5.10±0.01	2.53±2.23
UniGraph	1.18±0.03	1.39±0.02	5.04±0.02	2.54±2.17
UniCO	1.16±0.02	**1.31**±0.01	**3.77**±0.02	**2.08**±1.47

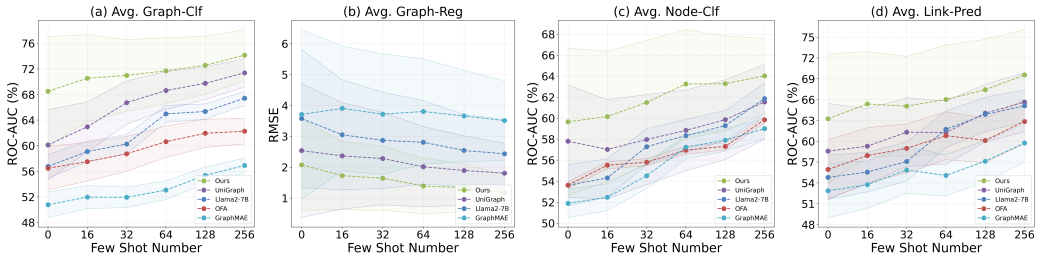


Figure 2: Performance trajectories of UniCO and baselines under few-shot fine-tuning.

details regarding datasets, data preprocessing, unified task formulation, pre-training implementations, and evaluation protocols, please refer to Appendix C.

5.2 Experimental Results and Analysis

5.2.1 Zero-Shot Generalization

Tables 1 to 3 report the zero-shot performance of UniCO and baseline models on unseen molecular graphs and text-attributed graphs. All downstream datasets are strictly excluded from pre-training, and model parameters remain completely frozen during inference.

As shown in Table 1, UniCO outperforms all baselines across the eight datasets, achieving an average ROC-AUC of 0.6849. This surpasses the specialized molecular baseline GIMLET and the cross-domain graph-text model UniGraph by 3.42% and 8.38%, respectively. Both the self-supervised graph model GraphMAE and the text-only LM LLaMA perform close to random guessing, indicating that relying solely on local message passing or linear text serialization fails to generalize to unseen graph-text scenarios. In Table 2, UniCO reduces the average RMSE to 2.08. On the `FreeSolv` benchmark, UniCO achieves a RMSE of 3.77, outperforming the second-best model by 1.27. This indicates that our covariance operators effectively mitigate numerical divergence caused by feature space shifts. In Table 3, UniCO achieves an average ROC-AUC of 0.5965 on node classification and 0.6323 on link prediction. Notably, on the `CiteSeer_edge` benchmark, UniCO attains a performance of 0.7324, exceeding all baselines by over 8%. These results confirm that our "text-graph" paradigm and unified covariance operators bridge multi-hop graph topologies with instruction contexts, enabling effective instruction-guided reasoning across domains and modalities.

5.2.2 Few-Shot Fine-Tuning

Figure 2 reports the performance of UniCO and representative baselines under few-shot fine-tuning with sample budgets $N_s \in \{0, 16, 32, 64, 128, 256\}$. The evaluation spans four downstream tasks.

As shown in Figures 2 (a) and 2 (c), UniCO establishes a clear advantage from its zero-shot initialization. With merely $N_s \leq 32$ labeled samples, it rivals the performance of baselines fine-tuned on 128 to 256 samples, while maintaining steady monotonic gains across all shot regimes. In Figure 2 (b), the RMSE of UniCO steadily declines as N_s increases from 0 to 64, mitigating prediction errors on continuous values in low-data regimes. Remarkably, in Figure 2 (d), UniCO requires only 64 labeled samples to surpass the 256-shot performance of all baselines. These results highlight UniCO's superior sample efficiency and rapid adaptation capabilities in low-resource scenarios. Detailed numerical results are provided in Appendix D.

5.2.3 Ablation Studies

To systematically verify the effectiveness of core components in UniCO, we report the zero-shot performance of five categories of ablation variants in Table 4.

As shown in Table 4, our ablation study examines five architectural designs: (1) Stochastic feature projections. Replacing the dynamically sampled Gaussian matrices with fixed Gaussian projections ($\mathbf{M}_g, \mathbf{M}_t$ fixed) leads to a marginal performance drop. Substituting them with conventional learnable fully connected layers ($\mathbf{R}^{(0)} = \mathrm{FC}(\mathbf{X})$) degrades generalization when encountering unseen input dimensions. (2) Covariance operator set. Completely removing the covariance opera-

Table 4: Zero-shot performance of UniCO and its ablation variants across representative benchmarks. Downstream tasks and evaluation metrics follow the setup in Sec. 5.2.1.

Task	Graph Classification		Graph Regression	Node Classification		Link Prediction
Model Variants	BACE	HIV	Lipo	Cora	WikiCS	CiteSeer_edge
$\mathbf{M}_g, \mathbf{M}_t$ fixed	$77.26_{\pm 0.22}$	$70.15_{\pm 0.25}$	$1.40_{\pm 0.03}$	$63.17_{\pm 0.38}$	$55.86_{\pm 0.41}$	$70.28_{\pm 0.33}$
$\mathbf{R}^{(0)} = \mathrm{FC}(\mathbf{X})$	$62.42_{\pm 0.30}$	$51.31_{\pm 0.35}$	$1.90_{\pm 0.02}$	$51.86_{\pm 0.27}$	$50.69_{\pm 0.34}$	$52.11_{\pm 0.37}$
w/o $\mathbf{C}$	$63.15_{\pm 0.31}$	$52.81_{\pm 0.36}$	$1.82_{\pm 0.01}$	$52.10_{\pm 0.28}$	$51.04_{\pm 0.35}$	$53.42_{\pm 0.39}$
w/o $\mathbf{C}_{tg}$	$68.24_{\pm 0.35}$	$59.45_{\pm 0.32}$	$1.63_{\pm 0.01}$	$55.32_{\pm 0.31}$	$52.49_{\pm 0.37}$	$56.85_{\pm 0.33}$
w/o $\mathbf{C}_{gg}, \mathbf{C}_{tt}$	$74.12_{\pm 0.30}$	$66.82_{\pm 0.28}$	$1.45_{\pm 0.02}$	$61.28_{\pm 0.37}$	$54.83_{\pm 0.40}$	$66.98_{\pm 0.21}$
$\mathbf{C}_{gt} \neq -\infty$	$73.58_{\pm 0.39}$	$65.13_{\pm 0.27}$	$1.51_{\pm 0.01}$	$59.84_{\pm 0.25}$	$53.98_{\pm 0.39}$	$65.11_{\pm 0.29}$
w/o $\mathbf{A}_{\mathrm{IF}}$	$76.87_{\pm 0.32}$	$69.56_{\pm 0.30}$	$1.41_{\pm 0.02}$	$62.75_{\pm 0.28}$	$55.41_{\pm 0.41}$	$68.86_{\pm 0.23}$
w/o $\mathbf{B}_t$	$78.15_{\pm 0.33}$	$71.02_{\pm 0.41}$	$1.36_{\pm 0.01}$	$63.93_{\pm 0.39}$	$56.26_{\pm 0.42}$	$70.95_{\pm 0.34}$
w/o $\mathcal{C}_{\mathrm{uni},e}$	$77.50_{\pm 0.28}$	$65.85_{\pm 0.37}$	$1.55_{\pm 0.02}$	–	–	–
$\mathbf{H}^{(0)} = \mathbf{S}$	$72.94_{\pm 0.28}$	$64.16_{\pm 0.26}$	$1.53_{\pm 0.02}$	$58.41_{\pm 0.24}$	$53.10_{\pm 0.38}$	$62.87_{\pm 0.27}$
$\mathbf{H}^{(0)} = \mathbf{R}^{(0)}$	$78.91_{\pm 0.34}$	$71.43_{\pm 0.31}$	$1.40_{\pm 0.01}$	$63.88_{\pm 0.29}$	$56.45_{\pm 0.42}$	$71.54_{\pm 0.34}$
UniCO	$\mathbf{80.86}_{\pm \mathbf{0.31}}$	$\mathbf{73.24}_{\pm \mathbf{0.22}}$	$\mathbf{1.31}_{\pm \mathbf{0.01}}$	$\mathbf{65.28}_{\pm \mathbf{0.27}}$	$\mathbf{57.29}_{\pm \mathbf{0.38}}$	$\mathbf{73.24}_{\pm \mathbf{0.30}}$

tors while retaining only self-loops and adjacency matrices (w/o $\mathbf{C}$) severely impairs performance across all tasks. Removing either the cross-modal covariance blocks (w/o $\mathbf{C}_{tg}$) or the intra-modal self-covariance blocks (w/o $\mathbf{C}_{gg}, \mathbf{C}_{tt}$) consistently causes performance degradation, validating the necessity of statistical co-variation in bridging cross-modal interactions. (3) Masking mechanism. Removing the unidirectional blocking mask to allow the graph to attend to instructions ($\mathbf{C}_{gt} \neq -\infty$) deteriorates performance across multiple benchmarks. Eliminating the information flow matrix (w/o $\mathbf{A}_{\mathrm{IF}}$) or the relative positional bias matrix (w/o $\mathbf{B}_t$) leads to noticeable performance drops, confirming the critical role of the masking mechanism. (4) Edge covariance operators. Removing the edge covariance operator set (w/o $\mathcal{C}_{\mathrm{uni},e}$) degrades performance on molecular benchmarks. (5) Initial hidden features. Relying solely on structural and positional encodings ($\mathbf{H}^{(0)} = \mathbf{S}$) or using only the projected features ($\mathbf{H}^{(0)} = \mathbf{R}^{(0)}$) underperforms the integrated formulation.

5.2.4 Sensitivity Analysis

In Figure 3, we evaluate the sensitivity of UniCO's zero-shot performance with respect to the maximum propagation order $K \in \{0, 1, 2, 3, 4, 5\}$ and the shared latent dimension $d_h \in \{32, 64, 128, 256, 512, 1024\}$ across representative benchmarks.

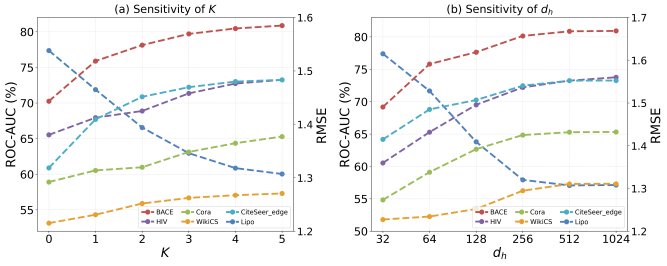


Figure 3: Sensitivity of UniCO's performance with respect to hyperparameters K and d_h.

As shown in Figure 3 (a), performance improves across benchmarks as K increases from 0 to 5. Since the operator set integrates self-loops, adjacency matrix, and multi-order operators, setting $K = 5$ matches the total operator count with the 8 attention heads of the backbone, enabling distinct heads to specialize in capturing multi-hop topological and contextual co-variations. In Figure 3 (b), performance is constrained under small latent dimensions ($d_h \leq 128$). As the dimension increases to $d_h = 512$ and 1024, the performance curve approaches its plateau. This confirms that a moderate latent dimension balances computational efficiency and predictive accuracy. Detailed analysis of computational complexity are provided in Appendix E.

6 Conclusion

In this paper, we propose UniCO, a graph foundation model framework built upon unified covariance operators to tackle cross-domain and cross-modal graph transferability. Specifically, we introduce a "text graph" paradigm that reduces the intricate dual "cross-modal and cross-domain" challenge into a singular cross-domain heterogeneous graph generalization problem. Furthermore, we construct unified covariance operators, incorporate an asymmetric masking mechanism, and mathematically establish that these operators reconcile "macroscopic modal decoupling" in expectation with "microscopic graph-text co-variation" while eliminating reliance on domain-specific feature spaces. Extensive experiments across diverse cross-domain graph tasks under zero-shot and few-shot settings demonstrate the superior transferability and generalization capabilities of UniCO.

REFERENCES

Haibo Chen, Xin Wang, Zeyang Zhang, Haoyang Li, Ling Feng, and Wenwu Zhu. Autogfm: Automated graph foundation model with adaptive architecture customization. In *Forty-second International Conference on Machine Learning*, 2025.

Jialin Chen, Haolan Zuo, Haoyu Wang, Siqi Miao, Pan Li, and Rex Ying. Towards a universal graph structural encoder. In *Proceedings of the ACM Web Conference 2026*, pp. 1457–1468, 2026.

Nuo Chen, Yuhan Li, Jianheng Tang, and Jia Li. Graphwiz: An instruction-following language model for graph computational problems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 353–364, 2024a.

Runjin Chen, Tong Zhao, Ajay Jaiswal, Neil Shah, and Zhangyang Wang. Llaga: Large language and graph assistant. *arXiv preprint arXiv:2402.08170*, 2024b.

Raffel Colin. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21, 2020.

Moshe Eliasof, Krishna Sri Ipsit Mantri, Beatrice Bevilacqua, Bruno Ribeiro, and Carola-Bibiane Schönlieb. Bridging input feature spaces towards graph foundation models. *arXiv preprint arXiv:2605.04834*, 2026.

Fabrizio Frasca, Fabian Jogl, Moshe Eliasof, Matan Ostrovsky, Carola-Bibiane Schönlieb, Thomas Gärtner, and Haggai Maron. Towards foundation models on graphs: An analysis on cross-dataset transfer of pretrained gnns. *arXiv preprint arXiv:2412.17609*, 2024.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

C Lee Giles, Kurt D Bollacker, and Steve Lawrence. Citeseer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pp. 89–98, 1998.

Haisong Gong, Zhibo Liu, Qiang Liu, Shu Wu, and Liang Wang. Nag: A unified native architecture for encoder-free text-graph modeling in language models. *arXiv preprint arXiv:2601.22657*, 2026.

Xiaoxin He, Xavier Bresson, Thomas Laurent, Bryan Hooi, et al. Explanations as features: Llm-based features for text-attributed graphs. *arXiv preprint arXiv:2305.19523*, 2(4):8, 2023.

Yufei He, Yuan Sui, Xiaoxin He, and Bryan Hooi. Unigraph: Learning a unified cross-domain foundation model for text-attributed graphs. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 448–459, 2025.

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 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 594–604, 2022.

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

Zihao Jing, Qiuhao Zeng, Ruiyi Fang, Yan Sun, Boyu Wang, and Pingzhao Hu. Entropy-guided dynamic tokens for graph-llm alignment in molecular understanding. *arXiv preprint arXiv:2602.02742*, 2026.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pp. 110–133, 2017.

Divyansha Lachi, Mehdi Azabou, Vinam Arora, and Eva Dyer. Graphfm: A scalable framework for multi-graph pretraining. *arXiv preprint arXiv:2407.11907*, 29, 2024.

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

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*, volume 2024, pp. 20188–20210, 2024.

Hugo Touvron Llama. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2407.21783*, 2023.

Haitao Mao, Zhikai Chen, Wenzhuo Tang, Jianan Zhao, Yao Ma, Tong Zhao, Neil Shah, Mikhail Galkin, and Jiliang Tang. Position: Graph foundation models are already here. *arXiv preprint arXiv:2402.02216*, 2024.

Péter Mernyei and Cătălina Cangea. Wiki-cs: A wikipedia-based benchmark for graph neural networks. *arXiv preprint arXiv:2007.02901*, 2020.

Bryan Perozzi, Bahare Fatemi, Dustin Zelle, Anton Tsitsulin, Mehran Kazemi, Rami Al-Rfou, and Jonathan Halcrow. Let your graph do the talking: Encoding structured data for llms, 2024. *URL https://arxiv. org/abs/2402.05862*, 1(3):4.

Kira Sam. Llama 3.1: An in-depth analysis of the next-generation large language model. *Available at SSRN 6139407*, 2024.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–106, 2008.

Yangyi Shen, Jincheng Zhou, Beatrice Bevilacqua, Joshua Robinson, Charilaos Kanatsoulis, Jure Leskovec, and Bruno Ribeiro. Zero-shot generalization of gnns over distinct attribute domains. 2025.

Sola Shirai, Kavitha Srinivas, Julian Dolby, Michael Katz, Horst Samulowitz, and Shirin Sohrabi. Less is more: Learning graph tasks with just llms. *arXiv preprint arXiv:2508.10115*, 2025.

Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*, 2022.

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

Li Sun, Zhenhao Huang, Silei Chen, Lanxu Yang, Junda Ye, Sen Su, and Philip Yu. Multi-domain riemannian graph gluing for building graph foundation models. In *International Conference on Learning Representations*, volume 2026, pp. 37129–37169, 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*, pp. 491–500, 2024.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

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

K Wang, Z Shen, C Huang, CH Wu, Y Dong, and A Kanakia. Microsoft academic graph: When experts are not enough. quantitative science studies, 1 (1), 396–413, 2020.

Shuo Wang, Bokui Wang, Zhixiang Shen, Boyan Deng, and Zhao Kang. Multi-domain graph foundation models: Robust knowledge transfer via topology alignment. *arXiv preprint arXiv:2502.02017*, 2025.

Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, Zhigang Zhou, Lianyi Han, Karen Karapetyan, Svetlana Dracheva, Benjamin A Shoemaker, et al. Pubchem's bioassay database. *Nucleic acids research*, 40(D1):D400–D412, 2012.

Zhihao Wen and Yuan Fang. 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*, pp. 506–516, 2023.

Jingyao Wu, Bin Lu, Zijun Di, Xiaoying Gan, Meng Jin, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. : One llm token for explicit graph structural understanding. In *International Conference on Learning Representations*, volume 2026, pp. 54029–54057, 2026.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Lianghao Xia and Chao Huang. Anygraph: Graph foundation model in the wild. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 882–896, 2026.

Lianghao Xia, Ben Kao, and Chao Huang. Opengraph: Towards open graph foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2365–2379, 2024.

Junjie Xu, Zongyu Wu, Minhua Lin, Xiang Zhang, and Suhang Wang. Llm and gnn are complementary: Distilling llm for multimodal graph learning. In *2025 IEEE international conference on big data (BigData)*, pp. 119–124. IEEE, 2025.

Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pp. 40–48. PMLR, 2016.

Xingtong Yu, Zechuan Gong, Chang Zhou, Yuan Fang, and Hui Zhang. Samgpt: Text-free graph foundation model for multi-domain pre-training and cross-domain adaptation. In *Proceedings of the ACM on Web Conference 2025*, pp. 1142–1153, 2025.

Haonan Yuan, Qingyun Sun, Junhua Shi, Xingcheng Fu, Bryan Hooi, Jianxin Li, and Philip S Yu. How much can transfer? bridge: Bounded multi-domain graph foundation model with generalization guarantees. In *Forty-second International Conference on Machine Learning*, 2025.

Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):862, 2022.

Haihong Zhao, Aochuan Chen, Xiangguo Sun, Hong Cheng, and Jia Li. All in one and one for all: A simple yet effective method towards cross-domain graph pretraining. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4443–4454, 2024.

Haiteng Zhao, Shengchao Liu, Ma Chang, Hannan Xu, Jie Fu, Zhihong Deng, Lingpeng Kong, and Qi Liu. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. *Advances in neural information processing systems*, 36:5850–5887, 2023.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*, 2020.

Yun Zhu, Haizhou Shi, Xiaotang Wang, Yongchao Liu, Yaoke Wang, Boci Peng, Chuntao Hong, and Siliang Tang. Graphclip: Enhancing transferability in graph foundation models for text-attributed graphs. In *Proceedings of the ACM on Web Conference 2025*, pp. 2183–2197, 2025.

A NOTATION LIST

The main mathematical notations and definitions are listed as follows.

Notations	Definition
$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A}_g)$	Input graph with node set $\mathcal{V}$, edge set $\mathcal{E}$, and adjacency matrix $\mathbf{A}_g$
$\lvert\mathcal{V}\rvert, \lvert\mathcal{E}\rvert$	Number of nodes and edges in graph $\mathcal{G}$
$\mathbf{X}_g \in \mathbb{R}^{\lvert\mathcal{V}\rvert \times d_n}$	Input node feature matrix of graph $\mathcal{G}$
d_n, d_e	Input node feature dimension and edge attribute dimension
$\mathbf{X}_e \in \mathbb{R}^{\lvert\mathcal{E}\rvert \times d_e}$	Input edge feature matrix of graph $\mathcal{G}$
$\boldsymbol{T}^\tau = (t_1, \ldots, t_m)$	Natural language instruction sequence for task τ with length m
$\mathbf{X}_t \in \mathbb{R}^{m \times d_t}$	Instruction token embedding matrix with dimension d_t
k_w	Sliding window size for constructing the "text graph"
$\mathbf{A}_t \in \mathbb{R}^{m \times m}$	Sliding-window adjacency matrix for the "text graph"
$N = \lvert\mathcal{V}\rvert + m$	Total sequence length of the unified joint graph
$\mathbf{A}_{\text{uni}} \in \mathbb{R}^{N \times N}$	Unified block-diagonal adjacency matrix $\text{diag}(\mathbf{A}_g, \mathbf{A}_t)$
$\boldsymbol{y}^\tau = (u_1, \ldots, u_S)$	Ground-truth target text sequence of length S
d_h	Shared latent feature projection dimension
$\mathbf{M}_g, \mathbf{M}_t$	Isotropic Gaussian random projection matrices for graph and text
$\mathbf{R}_g^{(0)}, \mathbf{R}_t^{(0)}$	Stochastic projected features of graph and text: $\mathbf{X}_g\mathbf{M}_g$ and $\mathbf{X}_t\mathbf{M}_t$
$\mathbf{\Pi}_g, \mathbf{\Pi}_t$	Geometric centering projection operators for graph and text
$\mathbf{R}_{g,c}^{(0)}, \mathbf{R}_{t,c}^{(0)}$	Centered initial projected features: $\mathbf{\Pi}_g\mathbf{R}_g^{(0)}$ and $\mathbf{\Pi}_t\mathbf{R}_t^{(0)}$
$\mathbf{R}_{\text{uni},c}^{(0)} \in \mathbb{R}^{N \times d_h}$	Unified centered initial feature matrix $[\mathbf{R}_{g,c}^{(0)\top}, \mathbf{R}_{t,c}^{(0)\top}]^\top$
K	Maximum propagation order of spatial diffusion
$\mathbf{R}_{\text{uni},c}^{(k)} \in \mathbb{R}^{N \times d_h}$	k-order diffused centered features
$\mathbf{C}_{\text{uni}}^{(k)} \in \mathbb{R}^{N \times N}$	k-order unified covariance operator $\frac{1}{d_h}\mathbf{R}_{\text{uni},c}^{(k)}(\mathbf{R}_{\text{uni},c}^{(k)})^\top$
$\mathbf{C}_{gg}^{(k)}, \mathbf{C}_{tt}^{(k)}$	Diagonal intra-modal covariance blocks of order k
$\mathbf{C}_{gt}^{(k)}, \mathbf{C}_{tg}^{(k)}$	Off-diagonal cross-modal covariance interaction blocks of order k
$\mathbf{A}_{\text{IF}} \in \mathbb{R}^{m \times m}$	Lower-triangular information flow matrix
$\mathbf{B}_t \in \mathbb{R}^{m \times m}$	Relative positional bias matrix from the LM backbone
$\mathbf{C}_{\text{uni-mask}}^{(k)}$	Asymmetrically modulated unified covariance operator
$\mathbf{I}_{\text{uni}}$	Unified self-loop matrix $\text{diag}(\mathbf{I}_g, \mathbf{I}_t)$
$\mathcal{C}_{\text{uni}}$	Set of unified covariance operators $\{\mathbf{I}_{\text{uni}}, \mathbf{A}_{\text{uni}}, \mathbf{C}_{\text{uni-mask}}^{(0)}, \ldots, \mathbf{C}_{\text{uni-mask}}^{(K)}\}$
$\mathbf{M}_e \in \mathbb{R}^{d_e \times d_h}$	Isotropic Gaussian random projection matrix for edge attributes
$\mathbf{R}_e^{(0)} \in \mathbb{R}^{\lvert\mathcal{V}\rvert \times d_h}$	Edge-induced initial node features
$\mathbf{C}_e^{(k)} \in \mathbb{R}^{\lvert\mathcal{V}\rvert \times \lvert\mathcal{V}\rvert}$	k-order edge-covariance operator
$\mathbf{C}_{\text{uni},e}^{(k)} \in \mathbb{R}^{N \times N}$	Unified edge covariance operator expanded with zero-padding
$\mathcal{C}_{\text{uni},e}$	Set of unified edge-covariance operators $\{\mathbf{0}_{\text{uni}}, \mathbf{0}_{\text{uni}}, \mathbf{C}_{\text{uni},e}^{(0)}, \ldots, \mathbf{C}_{\text{uni},e}^{(K)}\}$
$\mathcal{A}_{i,j}^{(l,k)}$	Operator-enhanced attention score between nodes i and j at head k and layer l
$\mathbf{W}_k^Q, \mathbf{W}_k^K$	Query and key projection matrices for the k-th attention head
α_k, β_k	Learnable scaling coefficients for unified node and edge covariance operators
$\mathbf{S} = \text{diag}(\mathbf{S}_g, \mathbf{S}_t)$	Integrated structural and positional encoding matrix
$\mathbf{H}^{(0)} \in \mathbb{R}^{N \times d_h}$	Initial hidden feature matrix: $\mathbf{R}_{\text{uni},c}^{(0)} \oplus \mathbf{S}$
L	Total number of Transformer encoder layers
$\mathbf{H}^{(L)} \in \mathbb{R}^{N \times d_h}$	Unified hidden feature matrix output by the final encoder layer
$\mathbf{H}_t^{(L)} \in \mathbb{R}^{m \times d_h}$	Text token hidden features extracted from $\mathbf{H}^{(L)}$ for cross-attention decoding
$\mathcal{T} = \{\tau\}$	Set of cross-domain graph prediction tasks
$\mathcal{L}_{\text{pretrain}}$	Sequence-length-normalized conditional autoregressive pre-training loss
Θ	Complete set of trainable model parameters

B Additional Theoretical Considerations and Proofs

B.1 Proofs for Geometric Preservation and Distributional Invariance

B.1.1 Foundational Lemmas on Gaussian Random Matrices

Lemma B.1.1. *Let $\mathbf{M} \in \mathbb{R}^{d\times d_h}$ be an isotropic Gaussian random matrix satisfying $\mathrm{vec}(\mathbf{M}) \sim \mathcal{N}(0, \mathbf{I}_{dd_h})$. Partitioning $\mathbf{M}$ into column vectors $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_{d_h}]$, the columns $\mathbf{m}_j \in \mathbb{R}^d$ are mutually independent and identically distributed standard multivariate normal vectors:*

$$\mathbf{m}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_d), \quad \forall j \in \{1, 2, \ldots, d_h\}. \tag{16}$$

Consequently, their first and second moments satisfy:

$$\mathbb{E}[\mathbf{m}_j] = \mathbf{0} \in \mathbb{R}^d, \quad \mathrm{Cov}(\mathbf{m}_j) = \mathbb{E}[(\mathbf{m}_j - \mathbb{E}[\mathbf{m}_j])(\mathbf{m}_j - \mathbb{E}[\mathbf{m}_j])^\top] = \mathbf{I}_d \in \mathbb{R}^{d\times d}. \tag{17}$$

B.1.2 Proof of Proposition 4.1.1 (Geometric Preservation)

Let $\mathbf{x}_u, \mathbf{x}_v \in \mathbb{R}^d$ ($d \in \{d_n, d_t\}$) denote an arbitrary pair of feature vectors corresponding to graph nodes or instruction tokens. Define the normalized difference vector as:

$$\mathbf{z} = \frac{\mathbf{x}_u - \mathbf{x}_v}{\|\mathbf{x}_u - \mathbf{x}_v\|_2} \in \mathbb{R}^{1\times d} \implies \|\mathbf{z}\|_2^2 = \sum_{i=1}^{d} z_i^2 = 1. \tag{18}$$

Consider the stochastic linear mapping $f(\mathbf{z}) = \frac{1}{\sqrt{d_h}}\mathbf{z}\mathbf{M} \in \mathbb{R}^{1\times d_h}$. Let $\mathbf{Y} = \mathbf{z}\mathbf{M} \in \mathbb{R}^{1\times d_h}$, whose j-th entry is given by:

$$Y_j = \sum_{i=1}^{d} z_i M_{i,j}. \tag{19}$$

By Lemma B.1.1, each entry $M_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Because independent Gaussian variables are closed under linear combinations, Y_j remains normally distributed. Its mean and variance evaluate to:

$$\mathbb{E}[Y_j] = \sum_{i=1}^{d} z_i \mathbb{E}[M_{i,j}] = 0, \tag{20}$$

$$\mathrm{Var}(Y_j) = \sum_{i=1}^{d} z_i^2 \mathrm{Var}(M_{i,j}) = \sum_{i=1}^{d} z_i^2 = 1. \tag{21}$$

Thus, $Y_1, \ldots, Y_{d_h} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. The squared norm of the projected vector satisfies:

$$\|f(\mathbf{z})\|_2^2 = \|\frac{1}{\sqrt{d_h}}\mathbf{z}\mathbf{M}\|_2^2 = \frac{1}{d_h}\sum_{j=1}^{d_h} Y_j^2 \implies d_h\|f(\mathbf{z})\|_2^2 \sim \chi^2_{d_h}, \tag{22}$$

which follows a standard chi-square distribution with d_h degrees of freedom.

Step 1: Moment-generating function of standard chi-square variables. Let $U \sim \mathcal{N}(0,1)$ with probability density function $f_U(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$. For any parameter $t < 1/2$, the moment-generating function (MGF) $M_{U^2}(t) = \mathbb{E}[e^{tU^2}]$ is derived via Gaussian integration:

$$\mathbb{E}[e^{tU^2}] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{tu^2 - \frac{u^2}{2}}du = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{1-2t}{2}u^2}du. \tag{23}$$

Letting $\alpha = \frac{1-2t}{2} > 0$, the Gaussian integral $I = \int_{-\infty}^{+\infty} e^{-\alpha u^2}du$ is evaluated by transforming into polar coordinates:

$$I^2 = (\int_{-\infty}^{+\infty} e^{-\alpha x^2}dx)(\int_{-\infty}^{+\infty} e^{-\alpha y^2}dy) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\alpha(x^2+y^2)}dxdy. \tag{24}$$

Substituting $x = r\cos\theta$, $y = r\sin\theta$, and $dxdy = rdrd\theta$:

$$I^2 = \int_0^{2\pi} d\theta \int_0^{+\infty} e^{-\alpha r^2} r dr = 2\pi \cdot [-\frac{1}{2\alpha} e^{-\alpha r^2}]_0^{+\infty} = \frac{\pi}{\alpha}, \tag{25}$$

which yields $I = \sqrt{\frac{\pi}{\alpha}} = \sqrt{\frac{2\pi}{1-2t}}$. Substituting I back into the expectation:

$$\mathbb{E}[e^{tU^2}] = \frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{2\pi}{1-2t}} = (1-2t)^{-1/2}, \quad \forall t < \frac{1}{2}. \tag{26}$$

Due to the mutual independence of $Y_1, \ldots, Y_{d_h}$, the joint MGF evaluates to:

$$\mathbb{E}[e^{t\sum_{j=1}^{d_h} Y_j^2}] = \prod_{j=1}^{d_h} \mathbb{E}[e^{tY_j^2}] = (1-2t)^{-d_h/2}. \tag{27}$$

Step 2: Chernoff bounding for tail probabilities. For any free parameter $t \in (0, 1/2)$, applying Markov's inequality yields the upper tail bound:

$$\mathbb{P}(\frac{1}{d_h}\sum_{j=1}^{d_h} Y_j^2 \geq 1+\epsilon) = \mathbb{P}(e^{t\sum Y_j^2} \geq e^{td_h(1+\epsilon)}) \leq \frac{\mathbb{E}[e^{t\sum Y_j^2}]}{e^{td_h(1+\epsilon)}} = (1-2t)^{-d_h/2} e^{-td_h(1+\epsilon)}. \tag{28}$$

To find the tightest bound, we minimize the exponent by defining $g(t) = -\frac{d_h}{2}\ln(1-2t) - td_h(1+\epsilon)$. Setting its first derivative to zero:

$$g'(t) = \frac{d_h}{1-2t} - d_h(1+\epsilon) = 0 \implies t^* = \frac{\epsilon}{2(1+\epsilon)} \in (0, \frac{1}{2}). \tag{29}$$

Evaluating $g(t^*)$ at the optimal point t^* gives:

$$g(t^*) = -\frac{d_h}{2}[\ln(\frac{1}{1+\epsilon}) + \frac{\epsilon}{1+\epsilon}(1+\epsilon)] = \frac{d_h}{2}[\ln(1+\epsilon) - \epsilon]. \tag{30}$$

Using the Taylor series expansion $\ln(1+\epsilon) \leq \epsilon - \frac{\epsilon^2}{2} + \frac{\epsilon^3}{3}$ for $\epsilon \in (0, 1/2)$, we bound $\ln(1+\epsilon) - \epsilon \leq -\frac{1}{2}(\epsilon^2 - \epsilon^3) \leq -\frac{\epsilon^2}{4}$. Substituting this back gives:

$$\mathbb{P}(\|f(\mathbf{z})\|_2^2 \geq 1+\epsilon) \leq \exp(-\frac{d_h\epsilon^2}{8}). \tag{31}$$

Applying a symmetric argument to the lower tail yields $\mathbb{P}(\|f(\mathbf{z})\|_2^2 \leq 1-\epsilon) \leq \exp(-\frac{d_h\epsilon^2}{8})$. Combining both tails via a union bound gives:

$$\mathbb{P}(|\|f(\mathbf{z})\|_2^2 - 1| \geq \epsilon) \leq 2\exp(-\frac{d_h\epsilon^2}{8}). \tag{32}$$

Step 3: Union bound on all node and token pairs. The total number of distinct pairs across all N tokens is $\binom{N}{2} < \frac{N^2}{2}$. Let $E_{u,v}$ denote the event that metric preservation fails for pair (u, v). Applying the union bound across all pairs:

$$\mathbb{P}(\exists u, v \in \mathcal{V} \cup \boldsymbol{T}^\tau,\ E_{u,v}) \leq \sum_{u \neq v} \mathbb{P}(E_{u,v}) < \frac{N^2}{2} \cdot 2\exp(-\frac{d_h\epsilon^2}{8}) = N^2\exp(-\frac{d_h\epsilon^2}{8}). \tag{33}$$

To guarantee that the joint failure probability is bounded by δ:

$$N^2\exp(-\frac{d_h\epsilon^2}{8}) \leq \delta \implies d_h \geq \frac{8\ln N^2 + 8\ln(1/\delta)}{\epsilon^2}. \tag{34}$$

Under this condition, the following holds with probability at least $1-\delta$:

$$(1-\epsilon)\|\mathbf{x}_u - \mathbf{x}_v\|_2^2 \leq \|f(\mathbf{x}_u) - f(\mathbf{x}_v)\|_2^2 \leq (1+\epsilon)\|\mathbf{x}_u - \mathbf{x}_v\|_2^2. \tag{35}$$

Step 4: Inner product preservation via polarization. For any two unit vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ ($\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$), the polarization identity in real inner product spaces states that:

$$\langle \mathbf{u}, \mathbf{v} \rangle = \frac{\|\mathbf{u}+\mathbf{v}\|_2^2 - \|\mathbf{u}-\mathbf{v}\|_2^2}{4}. \tag{36}$$

Since $f(\cdot)$ is linear, $f(\mathbf{u} \pm \mathbf{v}) = f(\mathbf{u}) \pm f(\mathbf{v})$. Substituting the metric preservation guarantees:

$$\langle f(\mathbf{u}), f(\mathbf{v}) \rangle = \frac{\|f(\mathbf{u}+\mathbf{v})\|_2^2 - \|f(\mathbf{u}-\mathbf{v})\|_2^2}{4} = \frac{\|\mathbf{u}+\mathbf{v}\|_2^2 - \|\mathbf{u}-\mathbf{v}\|_2^2}{4} \pm \frac{\epsilon}{4}(\|\mathbf{u}+\mathbf{v}\|_2^2 + \|\mathbf{u}-\mathbf{v}\|_2^2). \tag{37}$$

Applying the parallelogram identity $\|\mathbf{u}+\mathbf{v}\|_2^2 + \|\mathbf{u}-\mathbf{v}\|_2^2 = 2(\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2) = 4$:

$$\langle f(\mathbf{u}), f(\mathbf{v}) \rangle = \langle \mathbf{u}, \mathbf{v} \rangle \pm \frac{\epsilon}{4}(4) = \langle \mathbf{u}, \mathbf{v} \rangle \pm \epsilon, \tag{38}$$

which guarantees that with probability at least $1 - \delta$:

$$|\langle f(\mathbf{u}), f(\mathbf{v}) \rangle - \langle \mathbf{u}, \mathbf{v} \rangle| \leq \epsilon = \mathcal{O}(\epsilon). \tag{39}$$

B.1.3 Proof of Proposition 4.1.2 (Distributional Invariance)

Let $\mathbf{Q}_g \in \mathbb{R}^{d_n \times d_n}$ denote the orthogonal transformation matrix acting on the graph node feature space, satisfying $\mathbf{Q}_g \mathbf{Q}_g^\top = \mathbf{Q}_g^\top \mathbf{Q}_g = \mathbf{I}_{d_n}$. Define the auxiliary transformation matrix $\overline{\mathbf{M}}_g = \mathbf{Q}_g^\top \mathbf{M}_g \in \mathbb{R}^{d_n \times d_h}$. Partitioning $\overline{\mathbf{M}}_g$ into its column vectors yields $\overline{\mathbf{M}}_g = [\overline{\mathbf{m}}_1, \ldots, \overline{\mathbf{m}}_{d_h}]$, where each column is given by $\overline{\mathbf{m}}_j = \mathbf{Q}_g^\top \mathbf{m}_j$.

From Lemma B.1.1, we evaluate the first and second moments of $\overline{\mathbf{m}}_j$:

$$\mathbb{E}[\overline{\mathbf{m}}_j] = \mathbf{Q}_g^\top \mathbb{E}[\mathbf{m}_j] = \mathbf{Q}_g^\top \mathbf{0} = \mathbf{0}, \tag{40}$$

$$\mathrm{Cov}(\overline{\mathbf{m}}_j) = \mathbb{E}[(\mathbf{Q}_g^\top \mathbf{m}_j)(\mathbf{Q}_g^\top \mathbf{m}_j)^\top] = \mathbf{Q}_g^\top \mathbb{E}[\mathbf{m}_j \mathbf{m}_j^\top] \mathbf{Q}_g = \mathbf{Q}_g^\top \mathbf{I}_{d_n} \mathbf{Q}_g = \mathbf{I}_{d_n}. \tag{41}$$

Because an isotropic multivariate Gaussian distribution is invariant under orthogonal linear transformations, each column remains identically distributed as $\overline{\mathbf{m}}_j \sim \mathcal{N}(0, \mathbf{I}_{d_n})$. Furthermore, since the columns $\mathbf{m}_1, \ldots, \mathbf{m}_{d_h}$ are mutually independent, the transformed columns $\overline{\mathbf{m}}_1, \ldots, \overline{\mathbf{m}}_{d_h}$ remain mutually independent, which implies that $\overline{\mathbf{M}}_g$ and $\mathbf{M}_g$ are identically distributed:

$$\overline{\mathbf{M}}_g \overset{d}{=} \mathbf{M}_g. \tag{42}$$

For the transformed node features $\overline{\mathbf{X}}_g = \mathbf{X}_g \mathbf{Q}_g$, the resulting stochastic projected features satisfy:

$$\begin{aligned} \overline{\mathbf{R}}_g^{(0)} &= \overline{\mathbf{X}}_g \mathbf{M}_g = (\mathbf{X}_g \mathbf{Q}_g) \mathbf{M}_g \\ &\overset{d}{=} (\mathbf{X}_g \mathbf{Q}_g) \overline{\mathbf{M}}_g = (\mathbf{X}_g \mathbf{Q}_g)(\mathbf{Q}_g^\top \mathbf{M}_g) = \mathbf{X}_g (\mathbf{Q}_g \mathbf{Q}_g^\top) \mathbf{M}_g = \mathbf{X}_g \mathbf{M}_g = \mathbf{R}_g^{(0)}. \end{aligned} \tag{43}$$

Following an identical line of reasoning for the instruction token features, we obtain $\overline{\mathbf{R}}_t^{(0)} = (\mathbf{X}_t \mathbf{Q}_t) \mathbf{M}_t \overset{d}{=} \mathbf{R}_t^{(0)}$.

Because $\mathbf{M}_g$ and $\mathbf{M}_t$ are sampled independently, their concatenation preserves joint distributional equivalence after deterministic geometric centering:

$$\overline{\mathbf{R}}_{\text{uni},c}^{(0)} = \begin{bmatrix} \mathbf{\Pi}_g \overline{\mathbf{R}}_g^{(0)} \\ \mathbf{\Pi}_t \overline{\mathbf{R}}_t^{(0)} \end{bmatrix} \overset{d}{=} \begin{bmatrix} \mathbf{\Pi}_g \mathbf{R}_g^{(0)} \\ \mathbf{\Pi}_t \mathbf{R}_t^{(0)} \end{bmatrix} = \mathbf{R}_{\text{uni},c}^{(0)}. \tag{44}$$

Each k-th order covariance operator $\mathbf{C}_{\text{uni}}^{(k)} = \frac{1}{d_h} \mathbf{R}_{\text{uni},c}^{(k)} (\mathbf{R}_{\text{uni},c}^{(k)})^\top$ and its modulated counterpart $\mathbf{C}_{\text{uni-mask}}^{(k)}$ are computed as deterministic continuous functionals of $\mathbf{R}_{\text{uni},c}^{(0)}$. By the Continuous Mapping Theorem, applying deterministic functions to random variables that are equal in distribution yields outputs that are also equal in distribution. Consequently, the distributional equivalence directly extends to the joint distribution of the entire operator set:

$$\overline{\mathcal{C}}_{\text{uni}} \overset{d}{=} \mathcal{C}_{\text{uni}}. \tag{45}$$

B.2 Proofs for Expected Invariance and Macroscopic Modal Decoupling

B.2.1 Foundational Lemma on Outer Product Expectations

Lemma B.2.1. *Let $\mathbf{M} \in \mathbb{R}^{d\times d_h}$ be a Gaussian random matrix whose columns satisfy $\mathbf{m}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$ for $j \in \{1, \dots, d_h\}$. Then the mathematical expectation of the outer product matrix $\mathbf{M}\mathbf{M}^\top \in \mathbb{R}^{d\times d}$ evaluates to:*

$$\mathbb{E}_{\mathbf{M}}[\mathbf{M}\mathbf{M}^\top] = d_h \mathbf{I}_d. \tag{46}$$

Proof. Expanding $\mathbf{M}$ into the summation of its outer products $\mathbf{M}\mathbf{M}^\top = \sum_{j=1}^{d_h} \mathbf{m}_j \mathbf{m}_j^\top$. By the linearity of expectation:

$$\mathbb{E}_{\mathbf{M}}[\mathbf{M}\mathbf{M}^\top] = \sum_{j=1}^{d_h} \mathbb{E}[\mathbf{m}_j \mathbf{m}_j^\top]. \tag{47}$$

Using the identity $\mathrm{Cov}(\mathbf{m}_j) = \mathbb{E}[\mathbf{m}_j \mathbf{m}_j^\top] - \mathbb{E}[\mathbf{m}_j]\mathbb{E}[\mathbf{m}_j]^\top$, we have:

$$\mathbb{E}[\mathbf{m}_j \mathbf{m}_j^\top] = \mathrm{Cov}(\mathbf{m}_j) + \mathbb{E}[\mathbf{m}_j]\mathbb{E}[\mathbf{m}_j]^\top. \tag{48}$$

From Lemma B.1.1, $\mathbb{E}[\mathbf{m}_j] = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{m}_j) = \mathbf{I}_d$. Substituting these values yields $\mathbb{E}[\mathbf{m}_j \mathbf{m}_j^\top] = \mathbf{I}_d + \mathbf{0}\mathbf{0}^\top = \mathbf{I}_d$. Summing over all d_h independent channels gives:

$$\mathbb{E}_{\mathbf{M}}[\mathbf{M}\mathbf{M}^\top] = \sum_{j=1}^{d_h} \mathbf{I}_d = d_h \mathbf{I}_d. \tag{49}$$

B.2.2 Proof of Theorem 4.2.1 (Expected Invariance and Modal Decoupling)

Step 1: Annihilation of first-order statistics. Taking the graph modality as an exemplar, the centered projected features are given by $\mathbf{R}_{g,c}^{(0)} = \mathbf{\Pi}_g \mathbf{R}_g^{(0)} = (\mathbf{I}_{|\mathcal{V}|} - \frac{1}{|\mathcal{V}|}\mathbf{1}_{|\mathcal{V}|}\mathbf{1}_{|\mathcal{V}|}^\top)\mathbf{X}_g\mathbf{M}_g$. Computing the empirical sample mean along the node dimension:

$$\mathrm{Mean}(\mathbf{R}_{g,c}^{(0)}) = \frac{1}{|\mathcal{V}|}\mathbf{1}_{|\mathcal{V}|}^\top \mathbf{R}_{g,c}^{(0)} = \frac{1}{|\mathcal{V}|}\mathbf{1}_{|\mathcal{V}|}^\top (\mathbf{I}_{|\mathcal{V}|} - \frac{1}{|\mathcal{V}|}\mathbf{1}_{|\mathcal{V}|}\mathbf{1}_{|\mathcal{V}|}^\top)\mathbf{X}_g\mathbf{M}_g. \tag{50}$$

Expanding the matrix-vector product inside the parentheses:

$$\mathbf{1}_{|\mathcal{V}|}^\top (\mathbf{I}_{|\mathcal{V}|} - \frac{1}{|\mathcal{V}|}\mathbf{1}_{|\mathcal{V}|}\mathbf{1}_{|\mathcal{V}|}^\top) = \mathbf{1}_{|\mathcal{V}|}^\top - \frac{|\mathcal{V}|}{|\mathcal{V}|}\mathbf{1}_{|\mathcal{V}|}^\top = \mathbf{0}_{1\times|\mathcal{V}|}. \tag{51}$$

Hence, the empirical and expected first-order moments vanish identically across both modalities:

$$\mathrm{Mean}(\mathbf{R}_{g,c}^{(0)}) = \mathbf{0}_{1\times d_h}, \quad \mathrm{Mean}(\mathbf{R}_{t,c}^{(0)}) = \mathbf{0}_{1\times d_h}, \tag{52}$$

$$\mathbb{E}_{\mathbf{M}_g}[\mathbf{R}_{g,c}^{(0)}] = \mathbf{0}_{|\mathcal{V}|\times d_h}, \quad \mathbb{E}_{\mathbf{M}_t}[\mathbf{R}_{t,c}^{(0)}] = \mathbf{0}_{m\times d_h}. \tag{53}$$

Step 2: Derivation of covariance operator expectations and macro-level decoupling. The unified zero-order covariance operator $\mathbf{C}_{\text{uni}}^{(0)}$ is block-partitioned as:

$$\mathbf{C}_{\text{uni}}^{(0)} = \frac{1}{d_h}\begin{bmatrix} \mathbf{\Pi}_g\mathbf{X}_g\mathbf{M}_g\mathbf{M}_g^\top\mathbf{X}_g^\top\mathbf{\Pi}_g^\top & \mathbf{\Pi}_g\mathbf{X}_g\mathbf{M}_g\mathbf{M}_t^\top\mathbf{X}_t^\top\mathbf{\Pi}_t^\top \\ \mathbf{\Pi}_t\mathbf{X}_t\mathbf{M}_t\mathbf{M}_g^\top\mathbf{X}_g^\top\mathbf{\Pi}_g^\top & \mathbf{\Pi}_t\mathbf{X}_t\mathbf{M}_t\mathbf{M}_t^\top\mathbf{X}_t^\top\mathbf{\Pi}_t^\top \end{bmatrix}. \tag{54}$$

Since the centering matrices are symmetric ($\mathbf{\Pi}_g^\top = \mathbf{\Pi}_g, \mathbf{\Pi}_t^\top = \mathbf{\Pi}_t$), taking the expectation block-wise via Lemma B.2.1 yields for the diagonal blocks:

$$\mathbb{E}_{\mathbf{M}_g}[\mathbf{C}_{gg}^{(0)}] = \frac{1}{d_h}\mathbf{\Pi}_g\mathbf{X}_g\mathbb{E}_{\mathbf{M}_g}[\mathbf{M}_g\mathbf{M}_g^\top]\mathbf{X}_g^\top\mathbf{\Pi}_g = \frac{1}{d_h}\mathbf{\Pi}_g\mathbf{X}_g(d_h\mathbf{I}_{d_n})\mathbf{X}_g^\top\mathbf{\Pi}_g = \mathbf{\Pi}_g\mathbf{X}_g\mathbf{X}_g^\top\mathbf{\Pi}_g, \tag{55}$$

$$\mathbb{E}_{\mathbf{M}_t}[\mathbf{C}_{tt}^{(0)}] = \frac{1}{d_h}\mathbf{\Pi}_t\mathbf{X}_t\mathbb{E}_{\mathbf{M}_t}[\mathbf{M}_t\mathbf{M}_t^\top]\mathbf{X}_t^\top\mathbf{\Pi}_t = \mathbf{\Pi}_t\mathbf{X}_t\mathbf{X}_t^\top\mathbf{\Pi}_t. \tag{56}$$

Because $\mathbf{M}_g \in \mathbb{R}^{d_n \times d_h}$ and $\mathbf{M}_t \in \mathbb{R}^{d_t \times d_h}$ are drawn independently, $\mathbb{E}_{\mathbf{M}_g,\mathbf{M}_t}[\mathbf{M}_g\mathbf{M}_t^\top] = \mathbb{E}[\mathbf{M}_g]\mathbb{E}[\mathbf{M}_t]^\top = \mathbf{0}_{d_n \times d_t}$. Consequently, the cross-modal off-diagonal blocks vanish in expectation:

$$\mathbb{E}_{\mathbf{M}_g,\mathbf{M}_t}[\mathbf{C}_{gt}^{(0)}] = \frac{1}{d_h}\mathbf{\Pi}_g\mathbf{X}_g\mathbb{E}[\mathbf{M}_g\mathbf{M}_t^\top]\mathbf{X}_t^\top\mathbf{\Pi}_t = \mathbf{0}_{|\mathcal{V}|\times m}, \tag{57}$$

$$\mathbb{E}_{\mathbf{M}_g,\mathbf{M}_t}[\mathbf{C}_{tg}^{(0)}] = \frac{1}{d_h}\mathbf{\Pi}_t\mathbf{X}_t\mathbb{E}[\mathbf{M}_t\mathbf{M}_g^\top]\mathbf{X}_g^\top\mathbf{\Pi}_g = \mathbf{0}_{m\times|\mathcal{V}|}. \tag{58}$$

Step 3: Invariance to orthogonal transformations in expectation. Let $\overline{\mathbf{X}}_g = \mathbf{X}_g\mathbf{Q}_g$ and $\overline{\mathbf{X}}_t = \mathbf{X}_t\mathbf{Q}_t$ with orthogonal matrices satisfying $\mathbf{Q}_g\mathbf{Q}_g^\top = \mathbf{I}_{d_n}$ and $\mathbf{Q}_t\mathbf{Q}_t^\top = \mathbf{I}_{d_t}$. The expectation of the transformed graph operator evaluates to:

$$\begin{aligned}\mathbb{E}[\overline{\mathbf{C}}_{gg}^{(0)}] &= \frac{1}{d_h}\mathbf{\Pi}_g(\mathbf{X}_g\mathbf{Q}_g)\mathbb{E}[\mathbf{M}_g\mathbf{M}_g^\top](\mathbf{X}_g\mathbf{Q}_g)^\top\mathbf{\Pi}_g \\ &= \frac{1}{d_h}\mathbf{\Pi}_g\mathbf{X}_g\mathbf{Q}_g(d_h\mathbf{I}_{d_n})\mathbf{Q}_g^\top\mathbf{X}_g^\top\mathbf{\Pi}_g \\ &= \mathbf{\Pi}_g\mathbf{X}_g(\mathbf{Q}_g\mathbf{Q}_g^\top)\mathbf{X}_g^\top\mathbf{\Pi}_g = \mathbf{\Pi}_g\mathbf{X}_g\mathbf{X}_g^\top\mathbf{\Pi}_g = \mathbb{E}[\mathbf{C}_{gg}^{(0)}].\end{aligned} \tag{59}$$

Identical algebraic steps verify that $\mathbb{E}[\overline{\mathbf{C}}_{tt}^{(0)}] = \mathbb{E}[\mathbf{C}_{tt}^{(0)}]$. The off-diagonal zero blocks trivially preserve invariance under orthogonal transformations.

Step 4: Extension to multi-order operators. For any propagation order $k \geq 0$, the propagated centered features are given by $\mathbf{R}_{g,c}^{(k)} = \mathbf{\Pi}_g\mathbf{A}_g^k\mathbf{\Pi}_g\mathbf{X}_g\mathbf{M}_g$ and $\mathbf{R}_{t,c}^{(k)} = \mathbf{\Pi}_t\mathbf{A}_t^k\mathbf{\Pi}_t\mathbf{X}_t\mathbf{M}_t$. Applying Lemma B.2.1 directly:

$$\mathbb{E}_{\mathbf{M}_g}[\mathbf{C}_{gg}^{(k)}] = \mathbf{\Pi}_g\mathbf{A}_g^k\mathbf{\Pi}_g\mathbf{X}_g\mathbf{X}_g^\top\mathbf{\Pi}_g(\mathbf{A}_g^k)^\top\mathbf{\Pi}_g, \tag{60}$$

$$\mathbb{E}_{\mathbf{M}_t}[\mathbf{C}_{tt}^{(k)}] = \mathbf{\Pi}_t\mathbf{A}_t^k\mathbf{\Pi}_t\mathbf{X}_t\mathbf{X}_t^\top\mathbf{\Pi}_t(\mathbf{A}_t^k)^\top\mathbf{\Pi}_t. \tag{61}$$

Because $\mathbb{E}[\mathbf{M}_g\mathbf{M}_t^\top] = \mathbf{0}$, the cross-modal interaction blocks remain identically zero in expectation for any $k \geq 0$:

$$\mathbb{E}_{\mathbf{M}_g,\mathbf{M}_t}[\mathbf{C}_{gt}^{(k)}] = \mathbf{0}_{|\mathcal{V}|\times m}, \quad \mathbb{E}_{\mathbf{M}_g,\mathbf{M}_t}[\mathbf{C}_{tg}^{(k)}] = \mathbf{0}_{m\times|\mathcal{V}|}. \tag{62}$$

Since the masking matrix $\mathbf{A}_{\text{IF}}$, the relative positional bias $\mathbf{B}_t$, and the unilateral blocking mask are deterministic, the modulated operators $\mathbf{C}_{\text{uni-mask}}^{(k)}$ also preserve coordinate invariance in expectation.

B.2.3 Theoretical Conditions for Cross-Domain Transferability

Beyond orthogonal coordinate invariance, achieving successful transfer across graphs with fundamentally disparate feature spaces relies on the stability of the underlying structure captured by the expected operator $\mathbb{E}_{\mathbf{M}_g}[\mathbf{C}_{gg}^{(0)}] = \mathbf{\Pi}_g\mathbf{X}_g\mathbf{X}_g^\top\mathbf{\Pi}_g$. We posit that such stability arises when graphs across distinct domains share intrinsic structural and relational properties:

- **Homophily and Heterophily Invariants:** Graphs exhibit comparable relational structures tied to node features (e.g., connected or multi-hop reachable nodes share consistent co-variation patterns in their centered Gram matrices).
- **Structural Role Distributions:** Nodes with similar structural roles (e.g., hubs, bridges, or peripheral nodes) exhibit analogous feature activation patterns within their local neighborhoods across domains.
- **Shared Latent Generative Processes:** Graphs originate from related underlying generative mechanisms (such as shared Stochastic Block Models or Graphon formulations), yielding comparable spectral distributions in their centered Gram matrices $\mathbf{\Pi}_g\mathbf{X}_g\mathbf{X}_g^\top\mathbf{\Pi}_g$.

In these scenarios, even if the raw feature dimensionalities and realizations differ, the resulting covariance operators capture aligned relational geometries. Pre-training allows UniCO to learn domain-agnostic multi-order topological representations and instruction-following dynamics. Consequently, when transferring to unseen target domains, the backbone encoder can be kept completely frozen, requiring only a lightweight prediction head to achieve effective generalization.

B.3 Proofs for Structural Expressiveness and Spectral Boundedness

The algebraic definition of the multi-order cross-modal covariance operator $\mathbf{C}_{tg}^{(k)} \in \mathbb{R}^{m\times|\mathcal{V}|}$ is given by:

$$\mathbf{C}_{tg}^{(k)} = \frac{1}{d_h}\mathbf{R}_{t,c}^{(k)}(\mathbf{R}_{g,c}^{(k)})^\top = \frac{1}{d_h}(\mathbf{\Pi}_t\mathbf{A}_t^k\mathbf{\Pi}_t\mathbf{R}_t^{(0)})(\mathbf{\Pi}_g\mathbf{A}_g^k\mathbf{\Pi}_g\mathbf{R}_g^{(0)})^\top. \tag{63}$$

B.3.1 Proof of Proposition 4.3.1 (Structural Expressiveness)

Let $(\mathbf{R}_{t,c}^{(k)})_i \in \mathbb{R}^{1\times d_h}$ denote the i-th row of $\mathbf{R}_{t,c}^{(k)}$ (corresponding to the k-step propagated feature of instruction token i), and let $(\mathbf{R}_{g,c}^{(k)})_j \in \mathbb{R}^{1\times d_h}$ denote the j-th row of $\mathbf{R}_{g,c}^{(k)}$ (corresponding to the k-hop propagated feature of graph node j). By matrix multiplication, the (i,j)-th scalar entry of $\mathbf{C}_{tg}^{(k)}$ evaluates to the inner product of these two row vectors:

$$(\mathbf{C}_{tg}^{(k)})_{i,j} = \frac{1}{d_h}(\mathbf{R}_{t,c}^{(k)})_i(\mathbf{R}_{g,c}^{(k)})_j^\top. \tag{64}$$

Expanding each propagated row vector into its explicit linear combination over the initial centered features:

$$(\mathbf{R}_{t,c}^{(k)})_i = \sum_{p=1}^{m}(\mathbf{\Pi}_t\mathbf{A}_t^k\mathbf{\Pi}_t)_{i,p}(\mathbf{R}_{t,c}^{(0)})_p, \tag{65}$$

$$(\mathbf{R}_{g,c}^{(k)})_j = \sum_{q=1}^{|\mathcal{V}|}(\mathbf{\Pi}_g\mathbf{A}_g^k\mathbf{\Pi}_g)_{j,q}(\mathbf{R}_{g,c}^{(0)})_q. \tag{66}$$

Substituting these expansions into the inner product yields:

$$(\mathbf{C}_{tg}^{(k)})_{i,j} = \frac{1}{d_h}(\sum_{p=1}^{m}(\mathbf{\Pi}_t\mathbf{A}_t^k\mathbf{\Pi}_t)_{i,p}(\mathbf{R}_{t,c}^{(0)})_p)(\sum_{q=1}^{|\mathcal{V}|}(\mathbf{\Pi}_g\mathbf{A}_g^k\mathbf{\Pi}_g)_{j,q}(\mathbf{R}_{g,c}^{(0)})_q)^\top. \tag{67}$$

This algebraically confirms that each scalar entry $(\mathbf{C}_{tg}^{(k)})_{i,j}$ explicitly quantifies the empirical cross-covariance between the k-step contextual feature of instruction token i and the k-hop topological feature of graph node j.

B.3.2 Proof of Proposition 4.3.2 (Spectral Boundedness)

Applying the matrix-induced 2-norm (spectral norm) to $\mathbf{C}_{tg}^{(k)}$:

$$\begin{aligned}\|\mathbf{C}_{tg}^{(k)}\|_2 &= \frac{1}{d_h}\|(\mathbf{\Pi}_t\mathbf{A}_t^k\mathbf{\Pi}_t\mathbf{R}_t^{(0)})(\mathbf{\Pi}_g\mathbf{A}_g^k\mathbf{\Pi}_g\mathbf{R}_g^{(0)})^\top\|_2 \\ &= \frac{1}{d_h}\|\mathbf{\Pi}_t\mathbf{A}_t^k\mathbf{R}_{t,c}^{(0)}(\mathbf{R}_{g,c}^{(0)})^\top(\mathbf{A}_g^k)^\top\mathbf{\Pi}_g\|_2.\end{aligned} \tag{68}$$

By the sub-multiplicativity of induced matrix norms ($\|\mathbf{AB}\|_2 \le \|\mathbf{A}\|_2\|\mathbf{B}\|_2$):

$$\|\mathbf{C}_{tg}^{(k)}\|_2 \le \frac{1}{d_h}\|\mathbf{\Pi}_t\|_2\|\mathbf{A}_t^k\|_2\|\mathbf{R}_{t,c}^{(0)}\|_2\|(\mathbf{R}_{g,c}^{(0)})^\top\|_2\|(\mathbf{A}_g^k)^\top\|_2\|\mathbf{\Pi}_g\|_2. \tag{69}$$

We now bound each individual component:

(1) Centering projectors. Because $\mathbf{\Pi}_g$ and $\mathbf{\Pi}_t$ are orthogonal projection matrices satisfying $\mathbf{\Pi}^2 = \mathbf{\Pi} = \mathbf{\Pi}^\top$, their eigenvalues belong to $\{0,1\}$. Hence, their spectral norms satisfy exactly:

$$\|\mathbf{\Pi}_g\|_2 = 1, \quad \|\mathbf{\Pi}_t\|_2 = 1. \tag{70}$$

(2) Normalized adjacency matrices. The symmetrically normalized adjacency matrices $\mathbf{A}_g$ and $\mathbf{A}_t$ are real symmetric, with eigenvalues bounded in $[-1,1]$. Consequently, their spectral norms coincide with their spectral radii and are strictly bounded by 1:

$$\|\mathbf{A}_g\|_2 = \rho(\mathbf{A}_g) \le 1, \quad \|\mathbf{A}_t\|_2 = \rho(\mathbf{A}_t) \le 1. \tag{71}$$

By the sub-multiplicativity of norms, for any propagation order $k \geq 1$:

$$\|\mathbf{A}_g^k\|_2 \leq \|\mathbf{A}_g\|_2^k \leq 1^k = 1, \quad \|(\mathbf{A}_g^k)^\top\|_2 = \|\mathbf{A}_g^k\|_2 \leq 1, \quad \|\mathbf{A}_t^k\|_2 \leq \|\mathbf{A}_t\|_2^k \leq 1. \tag{72}$$

(2) Transposition invariance. Matrix transposition preserves the 2-norm: $\|(\mathbf{R}_{g,c}^{(0)})^\top\|_2 = \|\mathbf{R}_{g,c}^{(0)}\|_2$.

Substituting these bounds back into the inequality:

$$\begin{aligned}\|\mathbf{C}_{tg}^{(k)}\|_2 &\leq \frac{1}{d_h}(1) \cdot (1) \cdot \|\mathbf{R}_{t,c}^{(0)}\|_2 \cdot \|\mathbf{R}_{g,c}^{(0)}\|_2 \cdot (1) \cdot (1) \\ &= \frac{1}{d_h}\|\mathbf{R}_{t,c}^{(0)}\|_2 \|\mathbf{R}_{g,c}^{(0)}\|_2.\end{aligned} \tag{73}$$

This establishes that the spectral norm of $\mathbf{C}_{tg}^{(k)}$ is bounded by a constant that depends solely on the initial features, remaining strictly invariant to the propagation order k.

B.4 Proofs for Operator Consistency and Objective Upper Bound

B.4.1 Proof of Theorem 4.4.1 (Operator Convergence in Probability)

Let $\mathbf{W}_g = \mathbf{\Pi}_g \mathbf{X}_g \in \mathbb{R}^{|\mathcal{V}| \times d_n}$ denote the deterministic centered graph feature matrix. The empirical covariance operator $\mathbf{C}_{gg}^{(0)}$ can be expressed as the sample average of outer products over the column vectors $\mathbf{m}_1, \ldots, \mathbf{m}_{d_h}$ of the Gaussian matrix $\mathbf{M}_g \in \mathbb{R}^{d_n \times d_h}$:

$$\mathbf{C}_{gg}^{(0)} = \frac{1}{d_h}\mathbf{\Pi}_g \mathbf{X}_g \mathbf{M}_g \mathbf{M}_g^\top \mathbf{X}_g^\top \mathbf{\Pi}_g = \frac{1}{d_h}\mathbf{W}_g(\sum_{j=1}^{d_h} \mathbf{m}_j \mathbf{m}_j^\top)\mathbf{W}_g^\top = \frac{1}{d_h}\sum_{j=1}^{d_h}(\mathbf{W}_g \mathbf{m}_j)(\mathbf{W}_g \mathbf{m}_j)^\top. \tag{74}$$

Define the sequence of random matrices $\mathbf{Y}_j = (\mathbf{W}_g \mathbf{m}_j)(\mathbf{W}_g \mathbf{m}_j)^\top \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ for $j \in \{1, \ldots, d_h\}$.

By Lemma B.1.1, each column vector satisfies $\mathbf{m}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_{d_n})$. Because each $\mathbf{Y}_j$ is generated by an identical deterministic quadratic mapping of $\mathbf{m}_j$, the sequence $\mathbf{Y}_1, \ldots, \mathbf{Y}_{d_h}$ is independent and identically distributed. Applying Lemma B.2.1, its first moment evaluates to:

$$\mathbb{E}[\mathbf{Y}_j] = \mathbf{W}_g \mathbb{E}[\mathbf{m}_j \mathbf{m}_j^\top]\mathbf{W}_g^\top = \mathbf{W}_g \mathbf{I}_{d_n} \mathbf{W}_g^\top = \mathbf{\Pi}_g \mathbf{X}_g \mathbf{X}_g^\top \mathbf{\Pi}_g. \tag{75}$$

Now consider the scalar entry of $\mathbf{Y}_j$ at position (a, b):

$$(\mathbf{Y}_j)_{a,b} = (\mathbf{w}_a^\top \mathbf{m}_j)(\mathbf{w}_b^\top \mathbf{m}_j), \tag{76}$$

where $\mathbf{w}_a^\top, \mathbf{w}_b^\top \in \mathbb{R}^{1 \times d_n}$ denote the a-th and b-th rows of the constant matrix $\mathbf{W}_g$, respectively. Since the native node features are real and finite, the norms $\|\mathbf{w}_a\|_2$ and $\|\mathbf{w}_b\|_2$ are bounded. Standard Gaussian distributions possess finite moments of all orders; thus, the variance $\sigma_{ab}^2 = \mathrm{Var}((\mathbf{Y}_j)_{a,b}) < \infty$ exists and is strictly bounded.

The empirical entry $(\mathbf{C}_{gg}^{(0)})_{a,b} = \frac{1}{d_h}\sum_{j=1}^{d_h}(\mathbf{Y}_j)_{a,b}$ is the arithmetic mean of d_h i.i.d. scalar random variables with finite expectation and variance. By Khintchine's Weak Law of Large Numbers (WLLN), for any arbitrary $\eta > 0$:

$$\lim_{d_h \to \infty} \mathbb{P}(|(\mathbf{C}_{gg}^{(0)})_{a,b} - (\mathbf{\Pi}_g \mathbf{X}_g \mathbf{X}_g^\top \mathbf{\Pi}_g)_{a,b}| \geq \eta) = 0, \tag{77}$$

which establishes element-wise convergence in probability:

$$(\mathbf{C}_{gg}^{(0)})_{a,b} \xrightarrow{\mathcal{P}} (\mathbf{\Pi}_g \mathbf{X}_g \mathbf{X}_g^\top \mathbf{\Pi}_g)_{a,b} \quad \text{as } d_h \to \infty. \tag{78}$$

Because the graph contains a finite number of nodes $|\mathcal{V}| < \infty$, the squared Frobenius error comprises a finite sum of convergent terms:

$$\|\mathbf{C}_{gg}^{(0)} - \mathbf{\Pi}_g \mathbf{X}_g \mathbf{X}_g^\top \mathbf{\Pi}_g\|_F^2 = \sum_{a=1}^{|\mathcal{V}|}\sum_{b=1}^{|\mathcal{V}|}((\mathbf{C}_{gg}^{(0)})_{a,b} - (\mathbf{\Pi}_g \mathbf{X}_g \mathbf{X}_g^\top \mathbf{\Pi}_g)_{a,b})^2. \tag{79}$$

Since a finite sum of random variables converging in probability to zero also converges in probability to zero, we have:

$$\|\mathbf{C}_{gg}^{(0)} - \mathbf{\Pi}_g \mathbf{X}_g \mathbf{X}_g^\top \mathbf{\Pi}_g\|_F \xrightarrow{\mathcal{P}} 0 \quad \text{as } d_h \to \infty. \tag{80}$$

An identical derivation guarantees $\|\mathbf{C}_{tt}^{(0)} - \mathbf{\Pi}_t \mathbf{X}_t \mathbf{X}_t^\top \mathbf{\Pi}_t\|_F \xrightarrow{\mathcal{P}} 0$ for the instruction modality. For the cross-modal off-diagonal block $\mathbf{C}_{gt}^{(0)} = \frac{1}{d_h} \sum_{j=1}^{d_h} (\mathbf{\Pi}_g \mathbf{X}_g \mathbf{m}_{g,j})(\mathbf{\Pi}_t \mathbf{X}_t \mathbf{m}_{t,j})^\top$, the independence between $\mathbf{m}_{g,j}$ and $\mathbf{m}_{t,j}$ implies $\mathbb{E}[\mathbf{C}_{gt}^{(0)}] = \mathbf{0}$. Khintchine's WLLN directly confirms that $\mathbf{C}_{gt}^{(0)} \xrightarrow{\mathcal{P}} \mathbf{0}$ and $\mathbf{C}_{tg}^{(0)} \xrightarrow{\mathcal{P}} \mathbf{0}$. Hence, the full unified operator converges globally:

$$\mathbf{C}_{\text{uni}}^{(0)} \xrightarrow{\mathcal{P}} \mathbb{E}_{\mathbf{M}_g, \mathbf{M}_t}[\mathbf{C}_{\text{uni}}^{(0)}] = \begin{bmatrix} \mathbf{\Pi}_g \mathbf{X}_g \mathbf{X}_g^\top \mathbf{\Pi}_g & \mathbf{0}_{|\mathcal{V}| \times m} \\ \mathbf{0}_{m \times |\mathcal{V}|} & \mathbf{\Pi}_t \mathbf{X}_t \mathbf{X}_t^\top \mathbf{\Pi}_t \end{bmatrix} \quad \text{as } d_h \to \infty. \tag{81}$$

Let $\mathbf{\Sigma} = \mathbf{\Pi}_g \mathbf{X}_g \mathbf{X}_g^\top \mathbf{\Pi}_g$. The effective rank of $\mathbf{\Pi}_g \mathbf{X}_g$ is defined as:

$$r(\mathbf{\Pi}_g \mathbf{X}_g) = \frac{\text{Tr}(\mathbf{\Sigma})}{\|\mathbf{\Sigma}\|_2} = \frac{\|\mathbf{\Pi}_g \mathbf{X}_g\|_F^2}{\|\mathbf{\Pi}_g \mathbf{X}_g\|_2^2}. \tag{82}$$

According to the non-asymptotic concentration inequality for empirical covariance operators of sub-Gaussian vectors Koltchinskii & Lounici (2017), there exists an absolute universal constant $C > 0$ such that for any failure tolerance $\delta \in (0, 1)$, the relative spectral error satisfies with probability at least $1 - \delta$:

$$\frac{\|\mathbf{C}_{gg}^{(0)} - \mathbf{\Sigma}\|_2}{\|\mathbf{\Sigma}\|_2} \leq C \max\{\sqrt{\frac{r(\mathbf{\Pi}_g \mathbf{X}_g) + \ln(1/\delta)}{d_h}}, \frac{r(\mathbf{\Pi}_g \mathbf{X}_g) + \ln(1/\delta)}{d_h}\}. \tag{83}$$

Under the moderate over-sampling regime where $d_h \gtrsim r(\mathbf{\Pi}_g \mathbf{X}_g) + \ln(1/\delta)$, the first-order sub-Gaussian term dominates the quadratic sub-exponential tail:

$$\frac{r(\mathbf{\Pi}_g \mathbf{X}_g) + \ln(1/\delta)}{d_h} \leq \sqrt{\frac{r(\mathbf{\Pi}_g \mathbf{X}_g) + \ln(1/\delta)}{d_h}}. \tag{84}$$

Applying the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for non-negative numbers yields:

$$\frac{\|\mathbf{C}_{gg}^{(0)} - \mathbf{\Pi}_g \mathbf{X}_g \mathbf{X}_g^\top \mathbf{\Pi}_g\|_2}{\|\mathbf{\Pi}_g \mathbf{X}_g \mathbf{X}_g^\top \mathbf{\Pi}_g\|_2} \leq \mathcal{O}(\sqrt{\frac{r(\mathbf{\Pi}_g \mathbf{X}_g)}{d_h}} + \sqrt{\frac{\ln(1/\delta)}{d_h}}). \tag{85}$$

B.4.2 Proof of Proposition 4.4.2 (Training Objective Upper Bound)

By multivariate convex analysis, Jensen's inequality states that for any convex functional $f : \mathbb{R}^{m \times d_h} \to \mathbb{R}$ and any integrable random matrix $\mathbf{Z} \in \mathbb{R}^{m \times d_h}$ satisfying $\mathbb{E}[\|\mathbf{Z}\|] < \infty$:

$$f(\mathbb{E}[\mathbf{Z}]) \leq \mathbb{E}[f(\mathbf{Z})]. \tag{86}$$

In our framework, the stochasticity during each forward pass originates from the independently sampled Gaussian projection matrices $(\mathbf{M}_g, \mathbf{M}_t)$. Consequently, the final-layer text token representation $\mathbf{H}_t^{(L)}(\mathbf{M}_g, \mathbf{M}_t) \in \mathbb{R}^{m \times d_h}$ is a measurable multivariate random matrix with respect to $(\mathbf{M}_g, \mathbf{M}_t)$. For a target text sequence $\boldsymbol{y}_{\text{true}} = (u_1, u_2, \ldots, u_S)$, the autoregressive decoder evaluates the conditional log-likelihood via the standard cross-entropy loss:

$$\mathcal{L}(\mathbb{P}(\hat{\boldsymbol{y}} \mid \mathbf{H}_t^{(L)}), \boldsymbol{y}_{\text{true}}) = -\sum_{s=1}^{S} \log \mathbb{P}(u = u_s \mid u_{<s}, \mathbf{H}_t^{(L)}). \tag{87}$$

Define the composite objective function $f(\mathbf{H}_t^{(L)} = \mathcal{L}(\mathbb{P}(\hat{\boldsymbol{y}} \mid \mathbf{H}_t^{(L)}), \boldsymbol{y}_{\text{true}})$. In standard autoregressive sequence modeling, the composition of the softmax classification layer with the cross-entropy objective is globally convex with respect to the input unnormalized representations $\mathbf{H}_t^{(L)}$.

Substituting the random variable $\mathbf{H}_t^{(L)}(\mathbf{M}_g, \mathbf{M}_t)$ into Jensen's inequality:

$$f(\mathbb{E}_{\mathbf{M}_g, \mathbf{M}_t}[\mathbf{H}_t^{(L)}(\mathbf{M}_g, \mathbf{M}_t)]) \leq \mathbb{E}_{\mathbf{M}_g, \mathbf{M}_t}[f(\mathbf{H}_t^{(L)}(\mathbf{M}_g, \mathbf{M}_t))]. \tag{88}$$

Expanding both sides according to the definition of $f(\cdot)$ yields:

$$\mathcal{L}(\mathbb{P}(\hat{\boldsymbol{y}} \mid \mathbb{E}_{\mathbf{M}_g,\mathbf{M}_t}[\mathbf{H}_t^{(L)}]), \boldsymbol{y}_{\text{true}}) \leq \mathbb{E}_{\mathbf{M}_g,\mathbf{M}_t}[\mathcal{L}(\mathbb{P}(\hat{\boldsymbol{y}} \mid \mathbf{H}_t^{(L)}), \boldsymbol{y}_{\text{true}})]. \tag{89}$$

Here, the left-hand side represents the theoretical predictive loss evaluated on the expected stationary representation $\mathbb{E}[\mathbf{H}_t^{(L)}]$, while the right-hand side denotes the expected empirical optimization objective evaluated over dynamic forward projections. This demonstrates that minimizing the empirical stochastic loss directly enforces an upper bound on the true generalization risk of the stationary representation, ensuring stable optimization.

C Experimental Setup Details

C.1 Data Preprocessing and Unified Task Formulation

To systematically assess model generalizability across diverse domains, modalities, and task configurations, our benchmark spans two primary graph domains: molecular graphs and text-attributed graphs. We formulate four core downstream tasks, graph classification, graph regression, node classification, and link prediction, into a unified instruction-guided generative graph prediction paradigm, as defined in Sec. 3.1.

C.1.1 Molecular Graph Task Formulation

Molecular benchmarks primarily encompass graph-level bioactivity classification and continuous property regression. Following the standard evaluation protocol of GIMLET Zhao et al. (2023), we adopt the large-scale bioactivity repository `ChEMBL` Gaulton et al. (2012) as the source corpus for pre-training. For zero-shot downstream evaluation, we comprehensively evaluate on the MoleculeNet Wu et al. (2018) benchmark suite, covering bioactivity, physiology, toxicology, and physical chemistry properties (including `BACE`, `HIV`, `MUV`, `Tox21`, `ToxCast`, `BBBP`, `ESOL`, `Lipo`, and `FreeSolv`), alongside the large-scale `PCBA` Wang et al. (2012) benchmark and an independent held-out `ChEMBL Zero-Shot` (abbreviated as `ChEMBL-ZS`) test split.

C.1.2 Text-Attributed Graph Task Formulation

For large-scale text-attributed networks, to reconcile full-graph scalability with the compact graph scale of molecules, we reformulate node classification and link prediction into instruction-guided generative prediction over localized induced subgraphs.

Node Classification. For any target query node $v \in \mathcal{V}$, we extract its h-hop egocentric induced subgraph $\mathcal{G}_v = (\mathcal{V}_v, \mathcal{E}_v)$. We obtain the raw node text embedding matrix $\mathbf{X}_v \in \mathbb{R}^{|\mathcal{V}_v| \times d_n}$ and the localized adjacency matrix $\mathbf{A}_v \in \mathbb{R}^{|\mathcal{V}_v| \times |\mathcal{V}_v|}$. For datasets devoid of native edge attributes, the edge feature matrix is zero-padded to maintain dimensional compatibility across domains. We decouple the $|C|$-class categorization problem into $|C|$ independent binary instruction queries. To guarantee label balance during training and evaluation, for each positive node belonging to class C_i, we uniformly sample an equal number of negative nodes from the remaining classes, constructing a class-balanced evaluation set. The instruction $\boldsymbol{T}^\tau$ concatenates the target node's textual metadata (e.g., paper title and abstract) with a class-specific query (e.g., "`[Node Metadata] Does this node belong to class` $\{C_i\}$`?`"). The corresponding ground-truth target is formatted as a discrete token sequence $\boldsymbol{y}^\tau \in \{$"`Yes`", "`No`"$\}$.

Link Prediction. For any candidate node pair (u, v), we extract their respective $\lfloor h/2 \rfloor$-hop localized ego-subgraphs $\mathcal{G}_u = (\mathcal{V}_u, \mathcal{E}_u)$ and $\mathcal{G}_v = (\mathcal{V}_v, \mathcal{E}_v)$. Regardless of whether $\mathcal{G}_u$ and $\mathcal{G}_v$ overlap or connect in the original graph, we assemble them into a decoupled joint topology with a block-diagonal adjacency matrix $\mathbf{A}_{\text{pair}} = \text{diag}(\mathbf{A}_u, \mathbf{A}_v) \in \mathbb{R}^{(|\mathcal{V}_u|+|\mathcal{V}_v|)\times(|\mathcal{V}_u|+|\mathcal{V}_v|)}$ and the concatenated node feature matrix $\mathbf{X}_{\text{pair}} = [\mathbf{X}_u^\top, \mathbf{X}_v^\top]^\top \in \mathbb{R}^{(|\mathcal{V}_u|+|\mathcal{V}_v|)\times d_n}$. To alleviate the combinatorial explosion ($\mathcal{O}(|\mathcal{V}|^2)$) of candidate pairs on large-scale networks, we implement an unbiased evaluation protocol: across 5 independent random seeds, we sample 10% of the ground-truth edges as positive instances and uniformly sample an identical number of non-existent node pairs as negative instances, yielding a statistically sound evaluation benchmark. The instruction $\boldsymbol{T}^\tau$ integrates the metadata of both entities followed by a relational existence query (e.g., "`[Node` u `&` v `Metadata] Does`

Table 5: Summary statistics of pre-training and downstream evaluation benchmarks.

Splitting	Domain	Dataset	No. of Graphs / Nodes (Edges)	No. of Tasks	Task Type
Pre-training Datasets	Bioactivity Assay	ChEMBL Bioassay	365,065	1,048	Graph-Clf
	Physico-chemical	ChEMBL Property	365,065	13	Graph-Reg
	Academic	ogbn-ArXiv	169,343 (1,166,243)	40	Node-Clf
	Academic	ArXiv_2023	46,198 (78,543)	40	Node-Clf, Link-Pred
	Academic	PubMed	19,717 (44,338)	3	Node-Clf, Link-Pred
	Social	Reddit	33,434 (198,448)	1	Node-Clf, Link-Pred
Downstream Datasets	Large-Scale	PCBA	437,929	128	Graph-Clf
	Large-Scale	ChEMBL-ZS	91,266	262	Graph-Clf
	Pharmacokinetics	BBBP	2,039	1	Graph-Clf
	Bioactivity	MUV	93,087	17	Graph-Clf
	Bioactivity	BACE	1,513	1	Graph-Clf
	Bioactivity	HIV	41,127	1	Graph-Clf
	Toxicology	Tox21	7,831	12	Graph-Clf
	Toxicology	ToxCast	8,598	617	Graph-Clf
	Physico-chemical	ESOL	1,128	1	Graph-Reg
	Physico-chemical	FreeSolv	642	1	Graph-Reg
	Physico-chemical	Lipophilicity	4,200	1	Graph-Reg
	Academic	Cora	2,708 (5,429)	7	Node-Clf, Link-Pred
	Academic	CiteSeer	3,186 (4,277)	6	Node-Clf
	Web Knowledge	WikiCS	11,701 (215,863)	10	Node-Clf, Link-Pred
	Social	Instagram	11,339 (144,010)	1	Node-Clf, Link-Pred

a link exist between these two entities?"). The target sequence is labeled as "Yes" for connected pairs and "No" otherwise.

Pre-Training and Downstream Datasets. Pre-training is conducted on large-scale TAG repositories, including ogbn-ArXiv Wang et al. (2020), ArXiv_2023 He et al. (2023), PubMed Yang et al. (2016), and Reddit Huang et al. (2024). For zero-shot evaluation, the benchmark encompasses academic citation networks (Cora Sen et al. (2008), CiteSeer Giles et al. (1998)), a social network (Instagram Huang et al. (2024)), and a web encyclopedia knowledge graph (WikiCS Mernyei & Cangea (2020)).

C.1.3 Dataset Statistics

Table 5 summarizes the statistical characteristics of the pre-training and downstream evaluation datasets, spanning molecular graphs and text-attributed graphs across diverse domains.

C.1.4 Task Instruction Exemplars

To illustrate the unified instruction-guided generative formulation across heterogeneous tasks, we provide concrete input instruction exemplars for the four representative task types:

- Molecular Graph Classification (e.g., HIV-1 reverse transcriptase inhibition):

```
"HIV-1 RT is a multifunctional enzyme with RNA- and
DNA-dependent polymerase activities alongside ribonuclease
H (RNase H) activity, which degrades viral genomic RNA
templates during first-strand DNA synthesis to complete reverse
transcription and double-stranded DNA formation.  Is this
molecule active in this assay?"
```

- Molecular Graph Regression (e.g., Lipophilicity prediction):

```
"Lipophilicity is a critical physicochemical property of drug
candidates that modulates membrane permeability and solubility,
typically quantified by the octanol/water distribution
coefficient (log D at pH 7.4).  What is the octanol/water
distribution coefficient (log D at pH 7.4) of this molecule?"
```

- Node Classification on Academic Citation Networks (e.g., Cora paper topic classification):

```
"The node metadata is:  'Title:  Neural Networks with Quadratic
VC Dimension.  Abstract:  This paper shows that neural networks
which use continuous activation functions have VC dimension at
least as large as the square of the number of weights w.  This
result settles a long-standing open question, namely whether
the well-known O(w log w) bound, known for hard-threshold
```

```
nets, also held for more general sigmoidal nets.  Implications
for the number of samples needed for valid generalization are
discussed.'  Does the current node belong to Neural Networks?"
```

- Link Prediction on Social Networks (e.g., Instagram relation inference):

```
"The node metadata pair is:  'Node 1:  Follow us for rustic
home & holiday decor -- handmade & repurposed, thrifting trips,
flea markets, T-shirts, and more!  Grandview, TX' and 'Node 2:
The News- What It Means, Why It Matters.  Listen on AM 640 &
iHeartRadio app.  Los Angeles & Orange County News Links below'.
Does a link exist between these two entities?"
```

C.2 Pre-Training Implementation Details

Pre-training Scheme. We initialize our architecture using the pre-trained LM Colin (2020) backbone weights while freezing the base parameters, and partition all trainable parameters into two distinct optimization groups:

- LoRA Adaptation Group. We inject Low-Rank Adaptation (LoRA) modules across all self-attention, cross-attention, and feed-forward projection matrices of the T5 encoder and decoder (`q`, `k`, `v`, `o`, `wi`, `wi_0`, `wi_1`, and `wo`). We configure the rank to $r = 16$, the scaling factor to $\alpha_{\text{lora}} = 32$, and the dropout rate to 0.05. The down-projection matrices are initialized via standard Gaussian distributions, while the up-projection matrices are initialized to zero.
- Operator Parameter Group. This group comprises the learnable scaling coefficients α_k and β_k for the unified covariance operators, the relative positional bias matrix $\mathbf{B}_t$, and newly introduced LayerNorm parameters.

We deploy a phased cosine learning rate schedule for both groups. During the initial 10% of training steps (numerical calibration and warm-up), the LoRA group warms up from 5×10^{-5} to 1×10^{-4}, whereas the operator group starts from an elevated learning rate of 5×10^{-4} and decays toward 5×10^{-5} to rapidly stabilize operator initializations and prevent gradient instability. In the remaining 90% of steps, both parameter groups decay synchronously following a unified cosine annealing schedule. Validation ROC-AUC is monitored continuously; a min-heap structure dynamically maintains the top-10 checkpoints, facilitating seamless continual pre-training across domain extensions.

All experiments are conducted on a workstation equipped with an Intel Xeon Gold 6226R CPU, 256GB of RAM, an NVIDIA RTX 3090 GPU (24GB VRAM), running Ubuntu 18.04.6 LTS and PyTorch 1.13.0.

Unified Output Parsing and Decision Protocol. Following the decoding and vocabulary evaluation protocol of GIMLET, predictions across both classification and regression tasks are parsed directly from the autoregressive decoder outputs without external vocabulary masks:

- Classification Tasks: For graph classification, node classification, and link prediction, ground-truth targets are represented as binary text tokens `"Yes"` and `"No"`. During inference, the first generated token serves as the decision token. We extract the unnormalized logits corresponding to `"Yes"` and `"No"` at this initial decoding step and compute the normalized posterior probability via a binary softmax: $\mathbb{P}(\texttt{"Yes"}) = \frac{\exp(\text{Logit}(\texttt{"Yes"}))}{\exp(\text{Logit}(\texttt{"Yes"}))+\exp(\text{Logit}(\texttt{"No"}))}$. This continuous confidence score $\mathbb{P}(\texttt{"Yes"})$ is directly utilized to calculate the ROC-AUC across evaluation benchmarks. Guided by task instructions, the decoder naturally focuses probability mass on valid candidate tokens without constrained decoding heuristics.
- Regression Tasks: For molecular property regression, continuous targets are rounded to two decimal places and converted into floating-point strings. The decoder generates the numerical sequence autoregressively, including signs and decimal points. Following GIMLET, we apply the standard regular expression template `r"-?\d+\.?\d*e?-?\d*?"` to extract the generated numerical token string, parse it into a floating-point scalar as the predicted value, and compute the Root Mean Squared Error (RMSE).

C.3 Zero-Shot Evaluation and Few-Shot Fine-Tuning Details

Zero-Shot Evaluation Protocol. Zero-shot evaluation assesses the out-of-the-box generalization capability of UniCO when deployed to entirely unseen feature spaces, novel graph topologies, and unencountered task instructions. All downstream evaluation datasets are strictly excluded during pre-training, ensuring their feature distributions and graph topologies undergo no gradient updates.

For molecular graphs, we adopt scaffold splitting (8:1:1 train/validation/test ratio) based on core chemical topologies to evaluate structural out-of-distribution transfer. For text-attributed graphs, we apply standard random splitting (8:1:1 ratio). All zero-shot performance is evaluated strictly on the held-out test splits.

For multi-task benchmarks (e.g., `Tox21` comprising 12 toxicological endpoints, `Cora` comprising 7 node topic classification tasks), the model ingests the task-specific natural language instruction along with the corresponding graph instance, executing forward inference across individual subtasks independently. We report the mean ROC-AUC (for classification) or RMSE (for regression) averaged across all subtasks. Throughout zero-shot evaluation, all model parameters remain strictly frozen, loading the optimal pre-trained checkpoint without any test-time adaptation or gradient-based updates.

Few-Shot Fine-Tuning Protocol. For few-shot adaptation, we evaluate across support sample budgets $N_s \in \{16, 32, 64, 128, 256\}$. For classification tasks, we construct class-balanced support sets by uniformly sampling $N_s/|C|$ labeled instances per class from the training set. For regression tasks, we uniformly sample N_s labeled graphs. Validation metrics are monitored throughout fine-tuning, and we report the performance of the checkpoint that achieves the best validation score on the independent test set.

To mitigate catastrophic forgetting and severe overfitting in low-data regimes, we strictly constrain the trainable parameter footprint:

- Following the adaptation protocol of GIMLET, we unfreeze the final linear language modeling head projecting decoder states onto vocabulary logits, enabling rapid alignment of decision boundaries under low-resource supervision.
- We fine-tune the LoRA parameters injected into the attention modules alongside the learnable operator scaling coefficients α_k and β_k (constituting less than 2% of total model parameters), allowing the model to adapt its attention patterns for cross-modal interaction.

All native T5 Transformer layers and token embedding lookup tables remain completely frozen to preserve the underlying representations acquired during pre-training.

Fine-tuning is conducted using the Adam optimizer with a base learning rate of 5×10^{-4}. The batch size dynamically scales with the support budget N_s: when $N_s <$ batch_size, the batch size is proportionally reduced with gradient accumulation steps adjusted accordingly to align optimization dynamics. All few-shot results are averaged across multiple independent random seeds.

C.4 Baseline Models and Unified Evaluation Protocols

To ensure rigorous and equitable comparison, our benchmark encompasses 13 representative baselines spanning four major paradigms: self-supervised graph learning (GRACE Zhu et al. (2020), GraphMAE Hou et al. (2022)), graph topological prompt tuning (ZeroG Li et al. (2024), OFA Liu et al. (2024)), dual-encoder contrastive alignment (MoMu Su et al. (2022), G2P2 Wen & Fang (2023), GraphCLIP Zhu et al. (2025)), and multimodal generative graph foundation models (KV-PLM Zeng et al. (2022), Galactica-1.3B Taylor et al. (2022), GIMLET Zhao et al. (2023), GraphGPT Tang et al. (2024), LLaMA series Llama (2023); Touvron et al. (2023); Sam (2024), UniGraph He et al. (2025)). All baselines are benchmarked under identical data splits and task formulations.

C.4.1 Integration Mechanisms for Instructions and Graph Structures

Self-Supervised Graph Models (GRACE, GraphMAE): Operating as pure topology encoders without native text processing capabilities, these models incorporate instructions via offline textual em-

beddings. We encode natural language instructions into dense semantic vectors using a pre-trained Sentence-BERT model and introduce them into the graph topology as a global virtual super-node. Bidirectional links are constructed between the super-node and task-relevant target nodes (all nodes for graph-level tasks, the target node for node-level tasks, and the candidate node pair for link-level tasks), enabling cross-modal feature fusion via standard message passing.

Graph Topological Prompt Tuning Models (ZeroG, OFA): ZeroG encodes task instructions into semantic representations via a pre-trained language model, initializes them as prompt nodes, and fully connects them to all input graph nodes. OFA projects instruction features onto task prompt nodes while introducing explicit class prompt nodes that encode semantic descriptions of candidate classes. The task prompt, candidate class nodes, and input graph are unified into a single heterogeneous prompt graph.

Dual-Encoder Contrastive Alignment Models (MoMu, G2P2, GraphCLIP): Adopting a decoupled two-tower architecture, these frameworks process graph topologies through a graph encoder and task queries through an independent text encoder. The text encoder processes instructions concatenated with textual class descriptions into prompt sequences. Cross-modal contrastive objectives align textual class prototypes and graph representations within a shared metric space.

Multimodal Generative Models (KVPLM, Galactica-1.3B, GIMLET, GraphGPT, LLaMA series, UniGraph): KVPLM serializes molecular graphs into SMILES strings and appends a masked token placeholder ("`[Instruction Question] [MASK]`") under a masked language modeling (MLM) paradigm. Galactica-1.3B encodes SMILES strings surrounded by specialized molecular tags within a prompt template ("`Question: [Instruction] Answer:`"), directly generating output sequences autoregressively. GIMLET and LLaMA-based frameworks treat graph nodes as structured tokens, concatenating them with tokenized text sequences and injecting topological relational biases into the self-attention matrices. UniGraph encodes node/edge textual metadata via a text encoder, aggregates structural neighborhoods through Graph Attention Networks (GATs), and projects the resulting representations via a linear projection layer into continuous virtual soft tokens that prepended to LLM prompt templates. GraphGPT extracts graph representations using a pre-trained graph encoder and maps them via a two-layer MLP into a discrete sequence of graph tokens, replacing the `<graph>` placeholder in instruction templates for language model inference.

C.4.2 Pre-Training Configurations

All baseline models are pre-trained using their official open-source implementations and original optimization objectives, preserving their designated pre-training corpora and hyperparameters.

C.4.3 Downstream Evaluation Protocols

Self-Supervised Graph Models (GRACE, GraphMAE): A linear classification or regression head is appended to the frozen graph encoder. For zero-shot evaluation, predictions are derived via random projections or pre-trained node priors. For few-shot adaptation, the linear probing head is fine-tuned on the K-shot support set using cross-entropy loss (for classification) or mean squared error loss (for regression) while keeping the GNN encoder backbone completely frozen.

Graph Topological Prompt and Contrastive Models (ZeroG, OFA, MoMu, G2P2, GraphCLIP): OFA uses its unified classification head to compute logits over class prompt nodes, applying a softmax function to obtain posterior probabilities for zero-shot ROC-AUC evaluation. In the few-shot regime, the GNN backbone is frozen and only the output head is fine-tuned using cross-entropy loss (due to its structural formulation, OFA does not support continuous graph regression). ZeroG, G2P2, and GraphCLIP perform zero-shot classification by computing cosine similarities between graph representations and textual class prototypes without test-time updates; due to architectural constraints, they are not applicable to continuous property regression or few-shot tuning. MoMu evaluates zero-shot classification by measuring cosine similarities between molecular graph embeddings and the text embeddings of "`Yes`" and "`No`" prompts; for few-shot adaptation, only the final linear projection layer is fine-tuned.

Multimodal Generative Models (KVPLM, Galactica-1.3B, GIMLET, GraphGPT, LLaMA series, UniGraph): KVPLM is evaluated exclusively on classification tasks: in the zero-shot setting, ROC-AUC is computed from the normalized vocabulary probabilities of "`Yes`" and "`No`" at the `[MASK]`

token position; in few-shot settings, only the final language modeling head is fine-tuned (continuous property regression is unsupported due to MLM constraints). For Galactica-1.3B, GIMLET, GraphGPT, LLaMA series, and UniGraph, input sequence formatting and decision token parsing for zero-shot evaluation across graph classification, graph regression, node classification, and link prediction follow the unified generative protocol in Sec. C.2. For few-shot fine-tuning, only the final linear prediction head is updated, leaving all backbone weights frozen.

D Detailed Numerical Results for Few-Shot Fine-Tuning

To provide a comprehensive quantitative reference accompanying the few-shot adaptation trajectories shown in Figure 2, Table 6 details the downstream performance across varying few-shot budgets ($N_s \in \{16, 32, 64, 128, 256\}$) on graph-level tasks (Avg. Graph-Clf and Avg. Graph-Reg), while Table 7 presents the performance on node- and link-level tasks (Avg. Node-Clf and Avg. Link-Pred).

Table 6: Detailed downstream performance of UniCO and baseline models on graph-level tasks across varying few-shot budgets (16 to 256 shots).

Task	**Avg. Graph-Clf**					**Avg. Graph-Reg**				
Model	16-shot	32-shot	64-shot	128-shot	256-shot	16-shot	32-shot	64-shot	128-shot	256-shot
GraphMAE	$51.96_{\pm 1.79}$	$51.92_{\pm 1.61}$	$53.05_{\pm 1.45}$	$55.36_{\pm 1.30}$	$56.90_{\pm 1.17}$	$3.903_{\pm 2.02}$	$3.717_{\pm 1.95}$	$3.803_{\pm 1.70}$	$3.658_{\pm 1.48}$	$3.510_{\pm 1.29}$
OFA	$57.50_{\pm 3.02}$	$58.72_{\pm 2.72}$	$60.62_{\pm 2.45}$	$61.93_{\pm 2.21}$	$62.25_{\pm 1.99}$	–	–	–	–	–
Llama2-7B	$59.08_{\pm 1.64}$	$60.26_{\pm 1.48}$	$64.97_{\pm 1.33}$	$65.33_{\pm 1.20}$	$67.40_{\pm 1.08}$	$3.049_{\pm 1.78}$	$2.869_{\pm 1.56}$	$2.806_{\pm 1.35}$	$2.541_{\pm 1.17}$	$2.434_{\pm 1.11}$
UniGraph	$62.95_{\pm 3.89}$	$66.72_{\pm 3.40}$	$68.61_{\pm 2.96}$	$69.74_{\pm 2.56}$	$71.38_{\pm 2.20}$	$2.367_{\pm 1.72}$	$2.280_{\pm 1.50}$	$2.015_{\pm 1.30}$	$1.893_{\pm 1.12}$	$1.808_{\pm 0.97}$
UniCO	$\mathbf{70.54}_{\pm \mathbf{6.83}}$	$\mathbf{70.97}_{\pm \mathbf{5.60}}$	$\mathbf{71.70}_{\pm \mathbf{5.15}}$	$\mathbf{72.56}_{\pm \mathbf{4.54}}$	$\mathbf{74.13}_{\pm \mathbf{3.99}}$	$\mathbf{1.727}_{\pm \mathbf{1.10}}$	$\mathbf{1.643}_{\pm \mathbf{1.04}}$	$\mathbf{1.392}_{\pm \mathbf{0.90}}$	$\mathbf{1.335}_{\pm \mathbf{0.78}}$	$\mathbf{1.184}_{\pm \mathbf{0.57}}$

Table 7: Detailed downstream performance of UniCO and baseline models on node- and link-level tasks across varying few-shot budgets (16 to 256 shots).

Task	`Avg. Node-Clf`					`Avg. Link-Pred`				
Model	16-shot	32-shot	64-shot	128-shot	256-shot	16-shot	32-shot	64-shot	128-shot	256-shot
GraphMAE	$52.46_{\pm 1.28}$	$54.51_{\pm 1.20}$	$57.23_{\pm 1.12}$	$57.90_{\pm 1.05}$	$59.00_{\pm 0.98}$	$53.74_{\pm 3.42}$	$55.85_{\pm 3.38}$	$55.07_{\pm 2.96}$	$57.12_{\pm 2.94}$	$59.73_{\pm 2.74}$
OFA	$55.51_{\pm 0.33}$	$55.80_{\pm 0.35}$	$56.94_{\pm 0.38}$	$57.31_{\pm 0.34}$	$59.84_{\pm 0.31}$	$57.94_{\pm 4.01}$	$58.97_{\pm 3.52}$	$60.80_{\pm 3.46}$	$60.11_{\pm 3.22}$	$62.84_{\pm 3.17}$
Llama2-7B	$54.30_{\pm 1.82}$	$57.27_{\pm 1.72}$	$58.33_{\pm 1.57}$	$59.28_{\pm 1.44}$	$61.86_{\pm 1.45}$	$55.55_{\pm 2.85}$	$57.08_{\pm 2.82}$	$61.72_{\pm 2.65}$	$63.86_{\pm 2.48}$	$65.10_{\pm 2.30}$
UniGraph	$57.02_{\pm 4.73}$	$57.96_{\pm 4.29}$	$58.83_{\pm 3.85}$	$59.85_{\pm 3.80}$	$61.54_{\pm 3.56}$	$59.28_{\pm 5.43}$	$61.30_{\pm 4.95}$	$61.25_{\pm 4.55}$	$64.05_{\pm 4.15}$	$65.63_{\pm 4.25}$
UniCO	$\mathbf{60.15}_{\pm \mathbf{6.22}}$	$\mathbf{61.49}_{\pm \mathbf{5.91}}$	$\mathbf{63.25}_{\pm \mathbf{5.17}}$	$\mathbf{63.28}_{\pm \mathbf{4.54}}$	$\mathbf{64.02}_{\pm \mathbf{3.57}}$	$\mathbf{65.36}_{\pm \mathbf{7.57}}$	$\mathbf{65.05}_{\pm \mathbf{7.16}}$	$\mathbf{66.02}_{\pm \mathbf{7.85}}$	$\mathbf{67.41}_{\pm \mathbf{7.29}}$	$\mathbf{69.56}_{\pm \mathbf{6.51}}$

E Computational Complexity Analysis

To assess the scalability and computational feasibility of UniCO across large graph topologies and long instruction sequences, this section provides theoretical analyses of both time and space complexity, covering feature projection, multi-order diffusion, covariance operator construction, and operator-enhanced attention.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A}_g)$ denote the input graph with node feature matrix $\mathbf{X}_g \in \mathbb{R}^{|\mathcal{V}| \times d_n}$. Let the instruction comprise m tokens with embedding matrix $\mathbf{X}_t \in \mathbb{R}^{m \times d_t}$. The unified sequence length is $N = |\mathcal{V}| + m$. We denote the shared latent projection dimension by d_h, the maximum propagation order by K, and the number of Transformer encoder layers by L.

E.1 Time Complexity Analysis

The computational cost of a single forward pass decomposes into three principal stages:

Stage 1: Stochastic Feature Projection and Centering. Projecting raw node and text representations via isotropic Gaussian random matrices $\mathbf{M}_g \in \mathbb{R}^{d_n \times d_h}$ and $\mathbf{M}_t \in \mathbb{R}^{d_t \times d_h}$ incurs $\mathcal{O}(|\mathcal{V}| d_n d_h + m d_t d_h)$ time. Applying the geometric centering operators $\mathbf{\Pi}_g$ and $\mathbf{\Pi}_t$ requires row-wise mean subtractions, taking $\mathcal{O}(|\mathcal{V}| d_h + m d_h) = \mathcal{O}(N d_h)$ time. The complexity of this stage scales linearly with the input sizes:

$$T_{\text{proj}} = \mathcal{O}\big(|\mathcal{V}| d_n d_h + m d_t d_h\big). \tag{90}$$

Stage 2: Multi-Order Diffusion and Covariance Operator Construction. Spatial diffusion is performed over the block-diagonal unified adjacency matrix $\mathbf{A}_{\text{uni}} = \text{diag}(\mathbf{A}_g, \mathbf{A}_t)$ for orders $k \in \{0, \dots, K\}$. Via sparse matrix-matrix multiplication, graph propagation over $\mathbf{A}_g$ takes $\mathcal{O}(|\mathcal{E}|d_h)$ time per step. For the text graph, the sliding-window adjacency matrix $\mathbf{A}_t$ has a bounded node degree of $2k_w$, yielding a per-step propagation cost of $\mathcal{O}(mk_w d_h)$. Across all K orders, multi-order diffusion takes $\mathcal{O}(K(|\mathcal{E}| + mk_w)d_h)$ time. Explicitly computing the k-order covariance operators $\mathbf{C}_{\text{uni}}^{(k)} = \frac{1}{d_h}\mathbf{R}_{\text{uni},c}^{(k)}(\mathbf{R}_{\text{uni},c}^{(k)})^\top$ via outer products across K orders requires $\mathcal{O}(KN^2 d_h)$ operations. The asymmetric masking and relative positional bias additions incur $\mathcal{O}(K(m^2 + |\mathcal{V}|m))$ cost. Hence, the operator construction complexity is:

$$T_{\text{cov}} = \mathcal{O}\big(K(|\mathcal{E}| + mk_w)d_h + KN^2 d_h\big). \tag{91}$$

Stage 3: Operator-Enhanced Attention and Transformer Aggregation. For each Transformer encoder layer over sequence length N, computing standard scaled dot-product attention logits and applying softmax normalization requires $\mathcal{O}(N^2 d_h)$ time. Incorporating the unified covariance operators as structural attention biases requires only scalar additions, scaling as $\mathcal{O}(N^2)$. The position-wise feed-forward network incurs $\mathcal{O}(Nd_h^2)$ operations per layer. Across L stacked encoder layers, the computation time is:

$$T_{\text{trans}} = \mathcal{O}\big(L(N^2 d_h + Nd_h^2)\big). \tag{92}$$

Overall Time Complexity. Aggregating the three stages yields the total time complexity for the UniCO encoder:

$$T_{\text{total}} = \mathcal{O}\Big(|\mathcal{V}|d_n d_h + md_t d_h + K|\mathcal{E}|d_h + (K+L)N^2 d_h + LNd_h^2\Big). \tag{93}$$

In practical deployments, instructions are compact ($m \leq 512$), and node counts remain moderate under the induced subgraph formulation (averaging $|\mathcal{V}| \leq 50$). With fixed hidden dimension $d_h = 512$ and small propagation orders $K \leq 5$, the forward inference latency of UniCO is on the same order of magnitude as standard Graph Transformers.

Remark on Linear-Time Scalability for Full Graphs. When scaling to dense networks comprising tens of thousands to millions of nodes, explicit materialization of the $N \times N$ dense covariance operators $\mathbf{C}_{\text{uni}}^{(k)}$ can be bypassed entirely. By exploiting the associative property of matrix multiplication when projecting onto the attention value matrix $\mathbf{V} \in \mathbb{R}^{N \times d_h}$, one can compute $\mathbf{R}_{\text{uni},c}^{(k)}\big((\mathbf{R}_{\text{uni},c}^{(k)})^\top \mathbf{V}\big)$ instead of $\big(\mathbf{R}_{\text{uni},c}^{(k)}(\mathbf{R}_{\text{uni},c}^{(k)})^\top\big)\mathbf{V}$. Evaluating the intermediate inner product $(\mathbf{R}_{\text{uni},c}^{(k)})^\top \mathbf{V} \in \mathbb{R}^{d_h \times d_h}$ reduces the computational complexity from $\mathcal{O}(N^2 d_h)$ to $\mathcal{O}(Nd_h^2)$, enabling strictly linear scalability $\mathcal{O}(N)$ with respect to graph size.

E.2 Space Complexity Analysis

Parameter Footprint. The Gaussian projection matrices $\mathbf{M}_g$ and $\mathbf{M}_t$ are stochastically resampled at each forward pass and require no gradient caching. The multi-order covariance sets $\mathcal{C}_{\text{uni}}$ and $\mathcal{C}_{\text{uni},e}$ are derived via closed-form statistical operations and introduce zero trainable parameters. The newly introduced trainable parameters are restricted to the per-head scalar coefficients $\{\alpha_k, \beta_k\}$, layer normalizations, and low-rank LoRA matrices, which together constitute less than 2% of the total parameter budget.

Peak Activation Memory. Peak GPU memory allocation is primarily dominated by the projected feature matrices $\mathbf{R}_{\text{uni},c}^{(k)} \in \mathbb{R}^{N \times d_h}$ and the precomputed attention bias matrices $\mathbf{C}_{\text{uni-mask}}^{(k)} \in \mathbb{R}^{N \times N}$. Storing the intermediate diffusion features across K orders requires $\mathcal{O}(KNd_h)$ memory, while caching the attention bias matrices consumes $\mathcal{O}(KN^2)$ memory. With batch sizes of 96 or 128, training and inference execute reliably on workstations equipped with 24GB VRAM GPUs without experiencing out-of-memory bottlenecks.