面向图标记化大语言模型的重构式图指令调优

Zhongjian Zhang*
zhangzj@bupt.edu.cn 北京邮电大学 北京，中国

Xiao Wang*xiao_wang@buaa.edu.cn 北京航空航天大学 北京，中国

Mengmei Zhang
zhangmengmei@bestpay.com.cn 中国电信翼支付 北京，中国

Jiarui Tan
tanjiarui@bupt.edu.cn 北京邮电大学 北京，中国

Chuan Shi†
shichuan@bupt.edu.cn 北京邮电大学 北京，中国

arXiv:2603.01385v1 [cs.CL] 2 Mar 2026

摘要

大语言模型（LLM）的卓越成功促使研究人员将其应用为各种图相关任务的通用预测器，最终目标是开发能够泛化到多种场景的图基础模型。关键挑战在于将图数据与语言空间对齐，使大语言模型能够更好地理解图。作为一种受欢迎的范式，图标记化大语言模型（GTokenLLMs）将复杂结构和长文本编码为图标记序列，然后通过语言指令微调将其与文本标记对齐。尽管现有方法已取得初步成功，但我们的信息论分析表明，现有 GTokenLLMs 完全依赖语言指令提供的文本监督，只能实现隐式的图文对齐，从而产生文本主导偏差，未能充分利用图上下文。为克服这一局限，我们首先证明，对齐目标受限于输入图与其在大语言模型中隐藏表示之间的互信息上界，这促使我们提升该上界，以实现更好的对齐。为此，我们进一步提出一种重构式图指令微调流水线，**RGLM**[1]。我们的核心思想是从大语言模型的图令牌输出中重构图信息，显式引入图监督来约束对齐过程。在技术上，我们从两个互补视角出发，探索三种不同变体来实现 RGLM：输入空间中的 RGLM-Decoder，以及潜在空间中的 RGLM-Similarizer 和 RGLM-Denoiser。此外，我们还从理论上分析了各变体的对齐效果。在多个基准测试和任务场景上进行的大量实验验证了所提出 RGLM 的有效性，为 GTokenLLMs 的对齐研究开辟了新的方向。

CCS 概念

• **计算数学** → 图算法； • **计算方法** → 自然语言生成。

关键词

图基础模型；大语言模型；图挖掘

ACM 引用格式：
Zhongjian Zhang、Xiao Wang、Mengmei Zhang、Jiarui Tan 和 Chuan Shi。2026 年。通过重构式图指令微调迈向图标记化大语言模型。发表于《*2026* 年 *ACM* 万维网会议论文集》（*WWW '26*），*2026* 年 *4* 月 *13–17* 日，阿联酋迪拜。ACM，纽约，纽约州，美国，13 页。
https://doi.org/10.1145/3774904.3792077

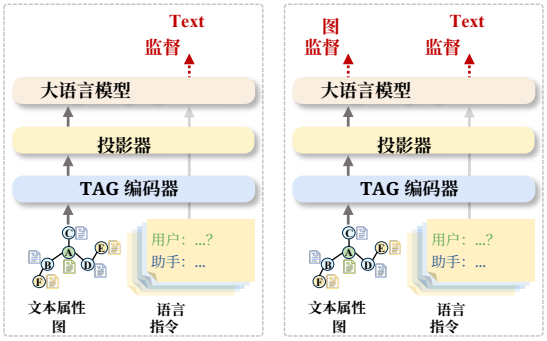
1 引言

图数据在现实世界中无处不在，例如社交网络 [18]和网页网络 [42]。传统图神经网络（GNN） [20, 44] 是建模图结构的强大工具，但其多任务处理能力较弱 [3]。近年来，GPT 风格大语言模型（LLM） [1, 6, 33] 的成功推动了众多研究将其应用于图学习领域，尤其是文本属性图（TAG），其中每个节点都关联着一个文本属性。最终目标是构建一种能够在多种场景中实现泛化的图基础模型 [25],；在这些场景中，大语言模型通常被用作预测器来完成各种图任务，例如节点分类，展现出良好的应用潜力并受到广泛关注 [5, 40, 56, 61]。

这一研究方向面临的一个核心挑战，是如何将图数据与自然语言嵌入空间对齐，使大语言模型能够更好地理解图。为应对这一挑战，现有工作主要分为两类：图文本化大语言模型（GTextLLMs） [5, 8, 10, 37, 59] 和图标记化大语言模型（GTokenLLMs） [3, 22, 32, 36, 55]。GTextLLMs 通过人工编写的提示模板，利用自然语言描述图，从而直接将其与文本对齐。然而，这类策略往往存在冗余和表征不准确的问题 [3, 5]。相比之下，GTokenLLMs 围绕图任务构建各种语言指令，以进行微调对齐，在保持训练方案经济高效的同时展现出更优的性能。如

*两位作者对本研究的贡献相同。†通讯作者。[1]源代码可在 https://github.com/zhongjian-zhang/RGLM 获取

本作品采用知识共享署名4.0国际许可协议授权。*WWW '26*，阿拉伯联合酋长国迪拜。© 2026 版权所有者/作者所有。ACM ISBN 979-8-4007-2307-0/2026/04 https://doi.org/10.1145/3774904.3792077



(a) 现有流程 (b) 我们的流水线

图1：（a）典型的图指令微调方法遵循以大语言模型为中心的设计，仅利用文本监督。（b）RGLM 设计了额外的图重构任务，将图监督显式纳入调优过程。

图1(a)所示，GTokenLLMs 通常采用 TAG 编码器，将复杂结构和长文本编码为图标记序列。随后，通过投影器将该序列与大语言模型嵌入空间对齐，并将其用作查询大语言模型的前缀令牌。然而，这一流水线：*TAG* 编码器 → 投影器→ 大语言模型 ⇐ 语言指令 （其中"⇐"表示监督），主要以大语言模型为中心：(i) 图理解在很大程度上取决于图文对齐和所选用的 TAG 编码器；(ii) 监督完全来源于文本数据。这自然引出了一个问题：仅凭语言指令能否有效支持图文对齐？回答这一问题对于推进更有效、更有原则的对齐策略至关重要。

在此，我们从信息论视角分析语言指令微调与图文对齐之间的关系。在 GTokenLLMs 中，理想的优化目标是：给定先前文本上下文，最大化图令牌对生成当前文本令牌的贡献，这可以通过最大化条件图-文本互信息（**C-GTMI**）来实现。我们证明，C-GTMI 可分解为两个主要项（见定理3.1）：（i）图文对齐项；（ii）文本自回归依赖项。然而，我们发现，当前的 GTokenLLMs [3, 32, 52] 主要优化第二项，却忽略了对第一项的显式优化，从而只能实现隐式图-文本对齐。这会导致模型偏向文本信息，可能削弱图上下文在图理解中的重要性，进而产生次优的 C-GTMI 和有限的对齐。因此，另一个问题自然产生：如何更有效地优化 *C-GTMI*，以实现更好的对齐？鉴于大语言模型的语义空间具有高维性，为此设计稳定且有效的优化机制仍然是一项重大挑战。

在本文中，我们首先证明 C-GTMI 存在一个上界
由输入图与其在大语言模型中的隐藏表示之间的互信息决定（见引理 3.1）。这一结果促使我们通过重构来改进该上界，从而实现更好的对齐。为此，我们进一步提出一种新颖的对齐流水线 **RGLM**，称为重构式图指令微调。我们的核心思路是从大语言模型的图令牌输出中重构图信息，而当前的 GTokenLLMs 在很大程度上忽略了这些输出。如

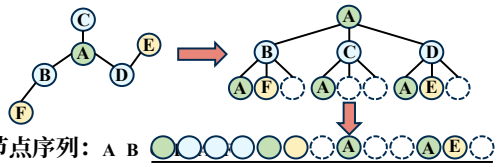


图 2：用于图序列化的邻居细节模板 [3]。

图1(b)所示，该流水线明确引入图监督，以约束对齐过程。具体而言，我们从两个互补的角度探索了三种变体，以实现 RGLM：从输入空间定义的 RGLM-Decoder，以及从潜在空间定义的 RGLM-Similarizer 和去噪器。此外，我们从理论上分析了每种 RGLM 变体如何优化 C-GTMI 的上界。我们的主要贡献如下：

- 我们首次从理论上揭示，当前采用语言指令微调的 GTokenLLMs 仅实现了隐式图-文本对齐，并证明对齐目标受输入图与其在大语言模型中隐表示之间的互信息所限制。
- 我们提出了一种重构式图指令微调流水线，称为 RGLM，并探索了三种不同变体：在输入空间中运行的 RGLM-Decoder，以及在潜在空间中运行的 RGLM-Similarizer 和 RGLM-Denoiser。此外，我们还从理论上分析了每种变体的对齐效果。
- 在四个数据集上的实验结果表明，所提出的 RGLM 通过引入不同的重构式图损失，在监督学习、多数据集泛化和零样本能力方面取得了显著提升，彰显了其作为图基础模型的潜力。

2 预备知识

2.1 文本属性图（TAGs）

我们将文本属性图定义为 $G_{\text{full}}=(V_{\text{full}}, E_{\text{full}}, S_{\text{full}})$，其中 $V_{\text{full}}=\{v_i\}_{i=1}^{|\mathcal{V}_{\text{full}}|}$、$E_{\text{full}}=\{e_i\}_{i=1}^{|\mathcal{E}_{\text{full}}|}$ 和 $S_{\text{full}}=\{s_i\}_{i=1}^{|\mathcal{S}_{\text{full}}|}$ 分别表示节点集、边集和文本集。文本集 S_{full} 通常由预训练语言模型 [7, 30]编码为节点特征 $Z_{\text{full}}=\{z1, \ldots, z|\mathcal{V}_{\text{full}}|\}$。为了使用大语言模型高效处理节点级任务（例如节点分类），通常会对以某个节点为中心的子图进行采样，以降低计算成本并减少无关信息 [3, 32]。形式化地，我们将以节点 v为中心的 h 跳子图表示为 $G_{\text{sub}}^{h}(v)=(V_v^h, E_v^h, Z_v^h)$，其中 V_v^h包含距离 v 不超过 h 跳的所有节点，E_v^h包含这些节点之间的边，Z_v^h则为对应的特征矩阵。为简洁起见，在上下文明确时，我们使用 $G=\{V, E, Z\}$ 表示此类子图。

2.2 图标记化大语言模型(GTokenLLMs)

一般而言，一个以 θ 参数化的大语言模型，会对长度为 T 的文本词元序列 $\boldsymbol{x}_{1:T}=(\boldsymbol{x}_1, \ldots\ldots, \boldsymbol{x}_T)$ 建模其典范因果分布 $p_\theta(\boldsymbol{x}) = \prod_{i=1}^{T} p_\theta(\boldsymbol{x}_i|\boldsymbol{x}_{<i})$。为了让大语言模型理解图内容，GTokenLLMs 会将复杂结构和长文本编码为图标记序列，并将其作为前缀令牌来查询大语言模型。如图 3(a) 所示，图 G 首先通过 TAG 编码器 F 编码为令牌序列，其中 F 可以是预定义的图神经网络 [14, 22, 55] ，也可以是图序列化模板 [3, 32, 52]。这些令牌随后被投影到大语言模型的

输入空间中，所用投影器 P 由 ϕ 参数化。因此，典范因果分布可重新表述为：

$$p_\Theta(\boldsymbol{x}) = \prod_{i=1}^{T} p_\Theta(\boldsymbol{x}_i|\boldsymbol{v}, \boldsymbol{x}_{<i}), \quad \boldsymbol{v} = \mathcal{P}_\phi \circ \mathcal{F}(\mathcal{G}), \tag{1}$$

其中，$\Theta = \{\theta, \phi\}$为参数，$\boldsymbol{v}_{1:N} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N)$是长度为 N的图标记序列。本文采用最先进的图序列化方法——邻居细节模板 [3] 作为 F，因为它对节点及其周围邻域进行了详尽描述，更适合我们后续的重构任务。如图 2所示，给定一个中心节点 A，我们构建一棵以 A 为根的固定大小计算树，用以表示其 h跳子图。在这棵树中，A 作为根节点，其 1 跳邻居被设为子节点，2 跳邻居被设为 1 跳邻居的子节点，以此类推。当节点的邻居数量超过预定义的大小时，我们随机采样其中一部分；若邻居数量不足，则插入占位节点以保持结构一致性。随后按层序遍历这棵树，生成固定长度的图标记序列，其中每个令牌对应一个节点，并以其特征$\mathbf{z}$表示。该过程能有效地将图结构转换为序列，同时保留层次化邻域信息。

通常，GTokenLLMs 通过有监督指令微调进行训练，模型参数 Θ 通过最小化语言监督（例如节点分类的文本标签）上的自回归损失进行优化。在这种设置下，自回归损失要求进行下一词元预测，即每个输入词元都会产生一个基于先前上下文的输出。总体目标是最大化有监督文本输出的对数似然。

$$\mathcal{L}_{\text{text}}(\Theta = \{\theta, \phi\}, \mathcal{G}, \boldsymbol{x}) = \frac{-1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \log p_\Theta(\boldsymbol{x}_i | \boldsymbol{v}, \boldsymbol{x}_{<i}), \tag{2}$$

其中，$\mathcal{S} \subseteq \{N+1, \ldots\ldots, T\}$表示输入序列中文本标签词元的索引集。如公式（2）所示，只有文本输出 $\boldsymbol{x}_{i \in \mathcal{S}}$ 会受到监督。

3 动机分析

3.1 现有 GTokenLLMs 的隐式图-文本对齐

给定一个长度为 T 的输入令牌序列，该序列由前缀图标记序列 $\boldsymbol{v}_{1:N} = (\boldsymbol{x}_1, \ldots\ldots, \boldsymbol{x}_N)$ 和文本词元序列 $\boldsymbol{x}_{1:M} = (\boldsymbol{x}_1, \ldots\ldots, \boldsymbol{x}_M)$ 组成，其中 $M = T - N$。在解码步 t，GTokenLLMs 基于 $\boldsymbol{v}$ 和 $\boldsymbol{x}_{<t}$ 预测 $\boldsymbol{x}_t$。我们用 $\mathrm{H}(\cdot)$ 表示熵，用 $\mathrm{I}(\cdot;\cdot)$ 表示互信息。GTokenLLMs 的理想优化目标是，在给定先前文本上下文的情况下，最大化图令牌对生成当前文本词元的贡献。为将其形式化，我们首先给出定义：

定义 3.1（条件图文互信息（C-GTMI））。在解码步 t，令 $\boldsymbol{s}^G$ 和 $\boldsymbol{s}^T_{<t}$ 分别表示大语言模型对图令牌和先前文本令牌的隐藏表示。*C-GTMI* 定义为 $\mathcal{I}(\boldsymbol{x}_t;\ \boldsymbol{s}^G \mid \boldsymbol{s}^T_{<t})$，用于衡量在给定先前文本上下文的情况下，图令牌对生成当前文本令牌的贡献程度。

基于上述定义，GTokenLLMs 的优化目标可以通过最大化 C-GTMI 来实现。

定理 3.1（通过自回归损失实现的隐式图-文本对齐）。*C-GTMI* 可以分解为：

$$\mathcal{I}(\boldsymbol{x}_t; \boldsymbol{s}^G \mid \boldsymbol{s}^T_{<t}) = \mathcal{I}(\boldsymbol{x}_t; \boldsymbol{s}^G) \ + \ \mathcal{I}(\boldsymbol{x}_t; \boldsymbol{s}^T_{<t} \mid \boldsymbol{s}^G) \ - \ \mathcal{I}(\boldsymbol{x}_t; \boldsymbol{s}^T_{<t}), \tag{3}$$

其中，$\mathcal{I}(\boldsymbol{x}_t;\ \boldsymbol{s}^G)$反映图文对齐，$\mathcal{I}(\boldsymbol{x}_t;\ \boldsymbol{s}^T_{<t} \mid \boldsymbol{s}^G)$捕捉文本自回归依赖，而 $\mathcal{I}(\boldsymbol{x}_t;\ \boldsymbol{s}^T_{<t})$仅取决于文本数据分布。

定理3.1表明，针对语言监督的自回归损失主要优化文本自回归依赖，但仅隐式地鼓励图文对齐，也就是说，$\mathrm{I}(\boldsymbol{x}_t;\ \boldsymbol{s}^G)$并未得到显式优化。因此，模型往往过度依赖文本信息，可能削弱图上下文在图理解中的重要性，导致 C-GTMI 不理想，性能受限。

证明。在实践中，GTokenLLMs 的自回归训练目标是真实分布 p与模型分布 q_θ之间的交叉熵损失。等价地，在步骤 t，这可以写为负对数似然：

$$\begin{aligned}\mathcal{L}_{\text{CE}} &= -\mathbb{E}\left[\log q_\theta(\boldsymbol{x}_t \mid \boldsymbol{v}, \boldsymbol{x}_{<t})\right] \\ &= \mathcal{H}(\boldsymbol{x}_t \mid \boldsymbol{v}, \boldsymbol{x}_{<t}) + \mathbb{E}\left[\mathrm{KL}p(\boldsymbol{x}_t \mid \cdot) \parallel q_\theta(\boldsymbol{x}_t \mid \cdot)\right] \\ &\geq \mathcal{H}(\boldsymbol{x}_t \mid \boldsymbol{v}, \boldsymbol{x}_{<t}).\end{aligned} \tag{4}$$

由于仅解码器大语言模型基于其内部隐藏状态而非直接基于原始输入词元生成词元，下一个文本词元的条件概率可以等价地用这些隐藏表示来表达。因此，我们有：

$$q_\theta(\boldsymbol{x}_t \mid \boldsymbol{v}, \boldsymbol{x}_{<t}) = q_\theta(\boldsymbol{x}_t \mid \boldsymbol{s}^G, \boldsymbol{s}^T_{<t}). \tag{5}$$

该表述强调，下一个词元的预测仅依赖于两种模态的内部隐藏表示，从而为在隐藏状态层面分析对齐提供了基础。将其代入公式（4）可得：

$$\mathcal{L}_{\text{CE}} = -\mathbb{E}[\log q_\theta(\boldsymbol{x}_t \mid \boldsymbol{s}^G, \boldsymbol{s}^T_{<t})] \geq \mathcal{H}(\boldsymbol{x}_t \mid \boldsymbol{s}^G, \boldsymbol{s}^T_{<t}). \tag{6}$$

根据互信息恒等式：

$$\mathcal{H}(\boldsymbol{x}_t \mid \boldsymbol{s}^G, \boldsymbol{s}^T_{<t}) = \mathcal{H}(\boldsymbol{x}_t \mid \boldsymbol{s}^T_{<t}) - \mathcal{I}(\boldsymbol{x}_t; \boldsymbol{s}^G \mid \boldsymbol{s}^T_{<t}). \tag{7}$$

直观而言，$\mathrm{H}(\boldsymbol{x}_t|\boldsymbol{s}^T_{<t})$ 由数据分布决定，最小化 $\mathrm{H}(\boldsymbol{x}_t|\boldsymbol{s}^G, \boldsymbol{s}^T_{<t})$ 等价于最大化 C-GTMI $\mathrm{I}(\boldsymbol{x}_t;\ \boldsymbol{s}^G \mid \boldsymbol{s}^T_{<t})$。根据互信息链式法则，我们有：

$$\mathcal{I}(\boldsymbol{x}_t; \boldsymbol{s}^G \mid \boldsymbol{s}^T_{<t}) = \mathcal{I}(\boldsymbol{x}_t; \boldsymbol{s}^G) \ + \ \mathcal{I}(\boldsymbol{x}_t; \boldsymbol{s}^T_{<t} \mid \boldsymbol{s}^G) \ - \ \mathcal{I}(\boldsymbol{x}_t; \boldsymbol{s}^T_{<t}). \tag{8}$$

□

3.2 迈向显式图–文本对齐

为通过更有效地优化 C-GTMI 来实现更好的对齐，我们首先从理论上建立 C-GTMI 的一个上界：

引理 3.1（C-GTMI 的上界）。给定一个输入图$\mathcal{G}$，*C-GTMI* 受图 $\mathcal{G}$ 与其在大语言模型中的隐藏表示 $\boldsymbol{s}^G$ 之间互信息的上界约束，即，

$$\mathcal{I}(\boldsymbol{x}_t; \boldsymbol{s}^G \mid \boldsymbol{s}^T_{<t}) \leq \mathcal{I}(\mathcal{G}; \boldsymbol{s}^G). \tag{9}$$

引理 3.1 促使我们优化 $\mathrm{I}(\mathrm{G};\ \boldsymbol{s}^G)$ 的下界，而这一改进可通过重构 [4, 15, 38, 39]，进一步实现，进而提高 C-GTMI 可达到的上界。这不仅显式优化了图文

对齐项 $\mathrm{I}(\boldsymbol{x}_t;\, \boldsymbol{s}^G)$，还增强了自回归依赖建模 $\mathrm{I}(\boldsymbol{x}_t;\, \boldsymbol{s}_{<t}^T \mid \boldsymbol{s}^G)$，从而实现更好的对齐。

证明。根据互信息链式法则，我们首先分析条件 $\boldsymbol{s}_{<t}^T$ 下的 $\mathrm{I}(\mathrm{G}\ \boldsymbol{s}^G,\ \boldsymbol{x}_t)$，得到：

$$\begin{aligned}\mathcal{I}(\mathcal{G};\boldsymbol{s}^G,\boldsymbol{x}_t \mid \boldsymbol{s}_{<t}^T) &= \mathcal{I}(\mathcal{G};\boldsymbol{s}^G \mid \boldsymbol{s}_{<t}^T) + \mathcal{I}(\mathcal{G};\boldsymbol{x}_t \mid \boldsymbol{s}^G,\boldsymbol{s}_{<t}^T)\\ &= \mathcal{I}(\mathcal{G};\boldsymbol{x}_t \mid \boldsymbol{s}_{<t}^T) + \mathcal{I}(\mathcal{G};\boldsymbol{s}^G \mid \boldsymbol{x}_t,\boldsymbol{s}_{<t}^T).\end{aligned} \tag{10}$$

将这两个表达式设为相等并重新排列各项，然后应用互信息对称性，可得：

$$\begin{aligned}&\mathcal{I}(\mathcal{G};\boldsymbol{s}^G \mid \boldsymbol{s}_{<t}^T) - \mathcal{I}(\mathcal{G};\boldsymbol{s}^G \mid \boldsymbol{x}_t,\boldsymbol{s}_{<t}^T)\\ &\quad= \mathcal{I}(\mathcal{G};\boldsymbol{x}_t \mid \boldsymbol{s}_{<t}^T) - \mathcal{I}(\mathcal{G};\boldsymbol{x}_t \mid \boldsymbol{s}^G,\boldsymbol{s}_{<t}^T)\\ &\quad= \mathcal{I}(\boldsymbol{x}_t;\mathcal{G} \mid \boldsymbol{s}_{<t}^T) - \mathcal{I}(\boldsymbol{x}_t;\mathcal{G} \mid \boldsymbol{s}^G,\boldsymbol{s}_{<t}^T)\\ &\quad= \mathcal{I}(\boldsymbol{x}_t;\boldsymbol{s}^G \mid \boldsymbol{s}_{<t}^T) - \mathcal{I}(\boldsymbol{x}_t;\boldsymbol{s}^G \mid \mathcal{G},\boldsymbol{s}_{<t}^T).\end{aligned} \tag{11}$$

鉴于图令牌 v 通常作为前缀令牌输入仅解码器大语言模型，因此 $\boldsymbol{s}^G$ 是由 G 确定性推导得出的，所以 $\mathrm{I}(\boldsymbol{x}_t;\, \boldsymbol{s}^G \mid \mathrm{G}\ \boldsymbol{s}_{<t}^T) = 0$，从而得到恒等式：

$$\begin{aligned}\mathcal{I}(\boldsymbol{x}_t;\boldsymbol{s}^G \mid \boldsymbol{s}_{<t}^T) &= \mathcal{I}(\mathcal{G};\boldsymbol{s}^G \mid \boldsymbol{s}_{<t}^T) - \mathcal{I}(\mathcal{G};\boldsymbol{s}^G \mid \boldsymbol{x}_t,\boldsymbol{s}_{<t}^T)\\ &= \mathcal{H}(\boldsymbol{s}^G \mid \boldsymbol{s}_{<t}^T) - \mathcal{I}(\mathcal{G};\boldsymbol{s}^G \mid \boldsymbol{x}_t,\boldsymbol{s}_{<t}^T)\\ &\le \mathcal{H}(\boldsymbol{s}^G \mid \boldsymbol{s}_{<t}^T).\end{aligned} \tag{12}$$

由于 $\mathrm{I}(\boldsymbol{s}^G;\, \boldsymbol{s}_{<t}^T) = \mathrm{H}(\boldsymbol{s}^G) - \mathrm{H}(\boldsymbol{s}^G \mid \boldsymbol{s}_{<t}^T) \ge 0$，因此可得：

$$\mathcal{H}(\boldsymbol{s}^G \mid \boldsymbol{s}_{<t}^T) \le \mathcal{H}(\boldsymbol{s}^G). \tag{13}$$

此外，由于 $\boldsymbol{s}^G$ 是 G 的确定性函数：

$$\mathcal{I}(\mathcal{G};\boldsymbol{s}^G) = \mathcal{H}(\boldsymbol{s}^G) - \mathcal{H}(\boldsymbol{s}^G \mid \mathcal{G}) = \mathcal{H}(\boldsymbol{s}^G). \tag{14}$$

合并这些不等式可得最终结果：

$$\mathcal{I}(\boldsymbol{x}_t;\boldsymbol{s}^G \mid \boldsymbol{s}_{<t}^T) \le \mathcal{I}(\mathcal{G};\boldsymbol{s}^G). \tag{15}$$

□

4 方法

4.1 概述

受引理 3.1 启发，我们提出了一种图重构指令微调流水线，称为 RGLM，通过额外的图重构任务显式优化 $\mathrm{I}(\mathrm{G};\, \boldsymbol{s}^G)$ 的下界。如图 3(a) 所示，首先使用邻居细节模板 F 将图 G 序列化为令牌序列。这些令牌通过由 ϕ 参数化的投影器与文本空间对齐，然后作为前缀令牌用于查询由 θ 参数化的大语言模型。我们的 RGLM 旨在从大语言模型生成的图令牌表示 $\boldsymbol{s}^G$ 中重构图信息。由于 $\boldsymbol{s}^G$ 可能包含表示同一节点的多个令牌（例如，图 2 中节点 A 出现了四次），我们对这些令牌进行均值聚合，以获得重构令牌表示 $\mathbf{H} = \{\mathbf{h}_1, \cdots, \mathbf{h}_{|\mathcal{V}|}\}$：

$$\mathbf{h}_v = \frac{1}{|\Gamma(v)|} \sum_{i \in \Gamma(v)} \boldsymbol{s}_i^G, \quad \forall v \in \mathcal{V}, \tag{16}$$

其中，$\Gamma(v) \subseteq \{1, \ldots\ldots, N\}$ 表示与 $\boldsymbol{s}^G$ 中节点 v 相关联的词元索引集。此处，为了降低计算成本并避免灾难性遗忘，[21], 我们采用 LoRA [16] 对大语言模型进行微调。训练目标由两个部分组成：

$$\mathcal{L}_{\mathrm{RGLM}} = \mathcal{L}_{\mathrm{text}} + \mathcal{L}_{\mathrm{graph}}, \tag{17}$$

其中，$\mathrm{L}_{\mathrm{text}}$ 是公式（2）中定义的文本自回归损失，$\mathrm{L}^{\mathrm{graph}}$ 是针对图数据的额外重建项。$\mathrm{L}^{\mathrm{graph}}$ 使用定制度量 M 定义，用于衡量重建令牌$\mathbf{H}$ 与从图 G 得出的预定义目标之间的差异。形式上，该损失表示为：

$$\mathcal{L}_{\mathrm{graph}}(\Theta = \{\phi, \theta, \pi\}, \mathcal{G}, \boldsymbol{x}) = \mathcal{M}\left(\mathcal{J}_\pi(\mathbf{H}), \mathcal{R}(\mathcal{G})\right), \tag{18}$$

其中，J_π 是一个由 π参数化的后投影函数，用于将$\mathbf{H}$映射为与预定义目标一致。函数 R 根据 G 构造重建目标，其中可以包含任何可观测的图信号。显然，R 和 M 的选择会产生不同的变体。接下来，我们从两个互补的角度探讨具有对齐保证的三种变体。在输入空间中，我们考虑一种直接重构原始节点特征和拓扑结构的变体。在预训练图神经网络的潜在空间中，我们考察两种重构潜在节点表示的语义变体：一种采用基于相似度的目标函数，另一种采用去噪策略。

4.2 输入空间中的原始图重构

在输入空间中，我们引入 RGLM-Decoder，从$\mathbf{H}$重构原始图信息。如图3(b)所示，我们训练一个特征解码器 d_f和一个结构解码器 d_s，分别重构节点特征和邻接关系。在此设置下，J_π 由 d_f和 d_s实例化，其中 R 由一个用于文本编码的预训练语言模型和一个邻接矩阵提取器组成。指标 M 定义为特征重构的均方误差（MSE）和结构重建的二元交叉熵（BCE）。目标函数表示为：

$$\mathcal{L}_{\mathrm{graph}} = \lambda_f \cdot \mathcal{L}_{\mathrm{feat}} + \lambda_s \cdot \mathcal{L}_{\mathrm{topo}}, \tag{19}$$

$$\mathcal{L}_{\mathrm{feat}} = \frac{1}{|\mathcal{V}|} \sum_{i=1}^{|\mathcal{V}|} \left\| d_f(\mathbf{h}_i) - \mathbf{z}_i \right\|_2^2, \tag{20}$$

$$\begin{aligned}\mathcal{L}_{\mathrm{topo}} = &-\frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \log \sigma\big(d_s(\mathbf{h}_i, \mathbf{h}_j)\big)\\ &- \frac{1}{|\hat{\mathcal{E}}|} \sum_{(i,j')\in\hat{\mathcal{E}}} \log\left(1 - \sigma(d_s(\mathbf{h}_i, \mathbf{h}_{j'}))\right),\end{aligned} \tag{21}$$

ˆ其中，E 和 E 分别表示 G 中已存在和不存在的连接集合，$\mathbf{z}$ 表示节点特征。$\sigma(\cdot)$ 是 Sigmoid函数，λ_f和 λ_s 分别用于平衡特征重构损失和结构重建损失。

命题4.1（通过RGLM-Decoder得到的下界）。给定一个图$\mathcal{G}$ =(Z,A)，其中Z 表示节点特征矩阵，A表示邻接矩阵，最小化重建损失$\mathcal{L}_{feat}$ 和$\mathcal{L}_{topo}$ 可为$\mathcal{I}(\mathcal{G};\, \boldsymbol{s}^G)$提供一个变分下界：

$$\mathcal{I}(\mathcal{G};\boldsymbol{s}^G) \ \ge\ \mathcal{H}(\mathbf{Z}) + \mathcal{H}(\mathbf{A}) \ -\ \tfrac{1}{\lambda_f}\mathcal{L}_{feat} \ -\ \tfrac{1}{\lambda_s}\mathcal{L}_{topo}. \tag{22}$$

证明见附录B.1。命题 4.1 表明，最小化原始图重建损失可最大化 $\mathrm{I}(\mathrm{G};\, \boldsymbol{s}^G)$ 的一个可计算下界，从而直接改进优化目标 C-GTMI $\mathrm{I}(\boldsymbol{x}_t;\, \boldsymbol{s}^G \mid \boldsymbol{s}_{<t}^T)$ 的上界。

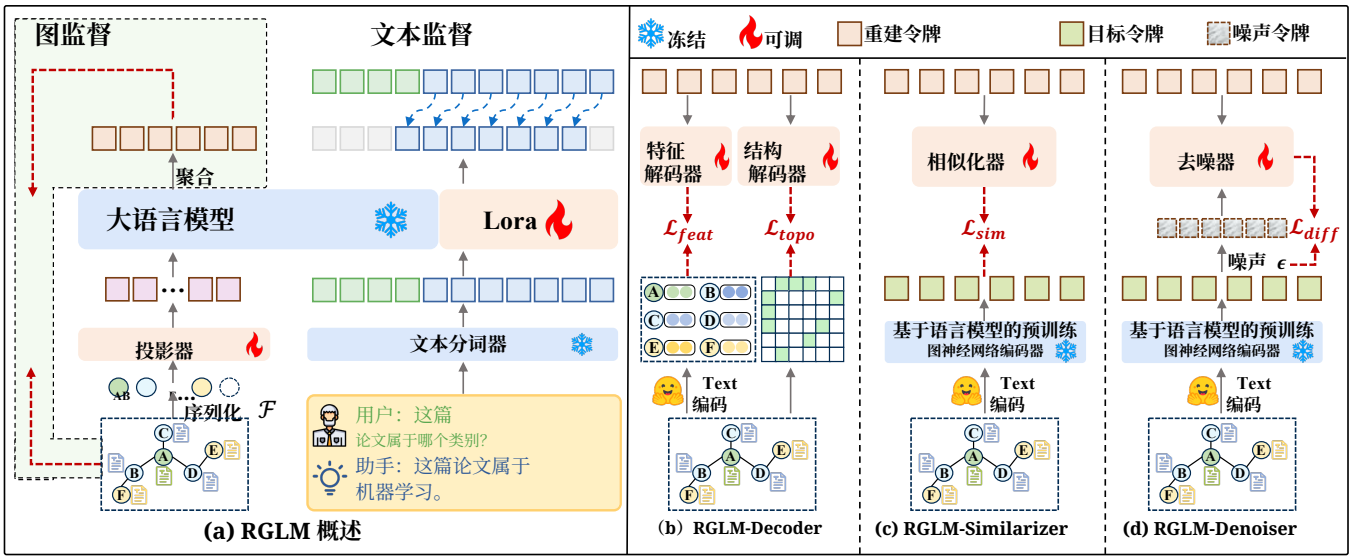


图3：(a) RGLM的概述。给定一个输入文本属性图，RGLM旨在从大语言模型的图令牌输出中重构图信息，并显式引入图监督以约束对齐过程。(b) 解码器变体直接在输入空间中重构原始节点特征与拓扑结构。(c) 相似化器和(d) 去噪器变体分别通过余弦相似度和去噪策略重构预训练图神经网络的潜在节点表示。

4.3 潜在空间中的节点表示重构

与在输入空间中重构原始图信息不同，本小节聚焦于在预训练图神经网络的潜在空间中重构节点表示。具体而言，我们首先预训练一个基于语言模型的图神经网络编码器 g 以获得细粒度的节点表示，然后训练一个轻量级模块——相似化器或去噪器——以从H中重构节点级语义。

基于语言模型的预训练***GNN***编码器。 为获得能够有效捕获离散图结构以及多层次特征信息（即节点级、边级和图级）的细粒度节点表示，我们预训练一个基于语言模型的图神经网络编码器 g，并将其置于自编码器框架 [50, 60]中。具体而言，给定完整文本属性图$\mathrm{G_{full}} = (\mathrm{V_{full}}, \mathrm{E_{full}}, \mathrm{S_{full}})$，我们首先使用 Sentence-BERT [30] 编码文本属性$\mathrm{S_{full}}$，以提取初始节点特征$\mathrm{Z_{full}}$。遵循 [29, 60], 以进一步增强结构感知，我们将缺失的边视为一种特殊边类型，从而引入可学习边嵌入，同时引入拉普拉斯位置编码和随机游走结构编码。我们采用DeeperGCN [23] 与图注意力网络 [35] 的组合作为编码器的骨干网络。由于潜在空间已经足够强大，解码器可以采用相对简单的线性层。在训练阶段，我们随机掩码训练集中的节点标签以进行预测，这可被视为图重构的一种特例。目标函数定义如下：

$$\mathcal{L}_{\text{pretrain}} = \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{reg}}, \tag{23}$$

其中，$\mathrm{L_{mask}}$ 表示掩码标签预测损失，$\mathrm{L^{reg}}$是一个正则化项，通过实现为 KL惩罚项，推动学习到的潜在表示趋向标准正态分布 [50]。编码器g训练充分后，将用于编码每个子图 G，并生成目标令牌表示E，其中g 对应公式（18）中的重构函数 R。

RGLM-Similarizer。 如图3(c)所示，RGLM-Similarizer 训练一个轻量级多层感知机 s，称为相似化器，直接将重建令牌表示H 回归到细粒度目标令牌表示E。在这种设定下，M 定义为余弦相似度损失，J 对应于 s。目标函数表示为：

$$\mathcal{L}_{\text{graph}} = \lambda_l \cdot \mathcal{L}_{\text{sim}}, \quad \mathcal{L}_{\text{sim}} = 1 - \frac{\langle s(\mathbf{H}), \mathbf{E} \rangle}{\|s(\mathbf{H})\|_2 \, \|\mathbf{E}\|_2}, \tag{24}$$

其中 λ_l 是用于平衡重建损失的缩放因子。

命题 4.2（通过 RGLM-Similarizer 得到的下界）。给定从预训练图神经网络 g中获得的图$\mathcal{G}$ 的潜在节点表示E，最小化余弦相似度损失$\mathcal{L}_{sim}$ 可为$\mathcal{I}(\mathcal{G};\ s^G)$提供一个变分下界：

$$\mathcal{I}(\mathcal{G}; s^G) \geq \mathcal{H}(\mathbf{E}) - (\kappa \cdot \lambda_l \cdot \mathcal{L}_{sim} + \mathrm{C}). \tag{25}$$

其中 $\kappa > 0$ 和C 是与 s^G无关的常数。证明见附录*B.2*。

RGLM-Denoiser。 RGLM-Similarizer 是一种结果导向的重构，它将H直接回归到E。相比之下，RGLM-Denoiser 利用去噪目标进行过程导向的重建。如图3(d)所示，RGLM-Denoiser 训练去噪器 f，以重建令牌H 为条件，从其噪声令牌$\mathbf{E}_t$中恢复干净潜在令牌E 。去噪器 f由一组Transformer块构成，每个块都配有三个额外的投影层，分别用于条件H、噪声输入$\mathbf{E}_t$和时间步 t。在这种设置下，M 定义为预测噪声与真实噪声之间的均方误差损失，J 对应于 f。训练过程遵循标准的扩散过程：

$$\mathcal{L}_{\text{graph}} = \lambda_l \cdot \mathcal{L}_{\text{diff}}, \quad \mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\epsilon}\left[\|f(\mathbf{E}_t, \mathbf{H}, t) - \epsilon\|_2^2\right], \tag{26}$$

表 1：监督设置下的节点分类结果。最佳结果以粗体标出，次佳结果以下划线标出。"内存溢出"表示内存不足，超过了 A800 的 80 GB 上限。"排名"表示在所有数据集上的综合排名。

方法	数据集	Cora		Pubmed		OGBN-Arxiv		Reddit		Rank	
	指标	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
GNNs	GCN	89.30	87.95	89.48	89.04	73.83	54.97	63.19	62.49	5	6
	GraphSAGE	87.82	86.77	90.54	90.51	74.47	56.41	58.51	58.41	10	8
	GAT	87.08	86.32	88.11	87.60	74.39	57.71	61.78	65.38	9	5
	DIFFormer	85.60	84.31	81.60	81.94	58.80	42.44	55.30	55.09	15	14
	UniMP	87.82	86.32	90.34	90.17	70.96	53.24	59.55	59.53	11	10
	NodeFormer	87.30	85.03	89.92	89.63	68.69	50.64	68.06	67.89	7	7
	GraphCL	86.53	84.50	90.39	90.28	OOM	OOM	57.67	57.67	12	12
	GRACE	89.11	88.12	86.89	86.58	72.00	50.79	65.39	65.39	8	9
	LGD	85.61	84.61	90.77	90.65	72.88	50.06	65.92	62.00	6	11
GTextLLMs	GraphText-ICL	82.29	81.16	56.72	52.55	68.67	51.67	56.50	52.23	16	16
GTokenLLMs	GraphText-SFT	84.69	83.77	87.80	87.68	49.88	19.46	62.20	62.07	14	15
	GraphGPT	83.92	82.72	81.76	80.99	71.59	54.58	62.11	62.39	13	13
	LLaGA	88.75	87.87	90.34	90.25	74.61	56.48	66.38	66.38	4	4
	RGLM-Decoder	89.85	**89.39**	91.15	91.07	75.00	56.92	**68.64**	**68.54**	1	1
	RGLM-Similarizer	89.67	88.96	**91.61**	**91.51**	**75.14**	**57.74**	66.79	66.77	3	2
	RGLM-Denoiser	**90.22**	89.35	90.95	90.90	75.10	56.84	67.64	67.50	2	3

其中 λ_l 是用于平衡重建损失的缩放因子。$\mathbf{E}_t$从 $\mathrm{N}(\sqrt{1-\beta_t}\mathbf{E}_{t-1}, \beta_t \mathbf{I})$ 中采样，而$\mathbf{E}_{t-1}$可以通过令 $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$ 直接从$\mathbf{E}$重新参数化：

$$\mathbf{E}_t = \sqrt{\bar{\alpha}_t}\mathbf{E} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{27}$$

命题 4.3（通过 RGLM-Denoiser 得到的下界）。给定由预训练图神经网络 g得到的图$\mathcal{G}$ 的潜在节点表示$\mathbf{E}$，最小化扩散噪声预测损失$\mathcal{L}_{diff}$可为$\mathcal{I}(\mathcal{G}; \boldsymbol{s}^G)$提供一个变分下界：

$$\mathcal{I}(\mathcal{G}; \boldsymbol{s}^G) \geq \mathcal{H}(\mathbf{E}) - (\lambda_l \cdot \mathcal{L}_{diff} + \mathrm{C}). \tag{28}$$

其中C 是一个与 $\boldsymbol{s}^G$无关的常数。证明见附录 *B.3*。

表2：监督设置下的链接预测结果。最佳结果以粗体标示，次优结果加下划线。

方法	数据集	Cora		Pubmed		Reddit	
	指标	Acc	F1	Acc	F1	Acc	F1
基线方法	GCN	84.26	83.91	90.00	89.98	74.38	73.85
	UniMP	71.62	70.92	80.76	80.54	73.09	72.53
	GRACE	70.44	69.19	81.74	81.56	61.68	60.88
	GraphGPT	59.85	59.85	73.60	73.39	56.69	27.19
	LLaGA	82.21	79.66	91.78	91.88	79.73	76.60
Ours	RGLM-Decoder	88.97	87.96	**94.84**	**94.91**	**81.90**	77.97
	RGLM-Similarizer	**89.85**	88.82	94.67	94.72	81.83	**78.10**
	RGLM-Denoiser	89.71	**88.92**	94.67	94.74	81.86	77.93

5 实验

5.1 实验设置

数据集。我们在四个广泛使用的数据集上进行实验：引文网络中的 Cora [51], Pubmed [51] 和 OGBN-Arxiv [17]，以及社交网络中的 Reddit [18]。遵循 [3,24],，我们选取两种广泛采用的指标：准确率（Acc(↑)）和宏平均 F1（F1(↑)）。更多数据集详情请参阅附录 C.1。

基线方法。我们比较了五类具有代表性的基线方法：(i) 传统图神经网络：GCN [20], GAT [35] 和 GraphSAGE [13]；(ii) 图 Transformer 模型：UniMP [31],DIFFormer [43]和 Node-Former [44]；(iii) 预训练-微调图神经网络：GraphCL [53],GRACE [62]和 LGD [60]；(iv) GTextLLMs：GraphText-ICL [59]；(v) GTokenLLMs：GraphText-SFT [59], GraphGPT [32] 和 LLaGA [3]。我们采用 Sentence-BERT [30] 从原始文本属性中提取节点特征，所有 GTokenLLMs 均使用 Vicuna-7B-v1.5-16K [6] 作为基础大语言模型。更多基线方法介绍和实现细节分别请参阅附录 C.2 和附录 C.3。

5.2 主要结果

5.2.1 节点分类性能。 我们主要关注有监督的节点分类任务。在表 1 中，我们发现：(1) RGLM 的表现始终优于所有基线方法。这凸显了 RGLM 涌现出的能力，使其能够处理超越传统图神经网络能力范围的复杂图任务。与现有的 GTokenLLMs 相比，RGLM 引入了显式的图监督，从而实现更有效的对齐，带来更优的性能。(2) 在 RGLM 的三个变体中，Decoder 的整体表现最佳。一个可能的原因是，在输入空间中同时重构节点特征和图拓扑结构，能够为对齐提供更全面的监督。(3) 尽管我们遵循 LGD 的预训练框架来学习节点潜在表示，但 Similarizer/Denoiser 的性能并未受到其上限的限制。这是因为图重构充当正则化项，旨在提升模型可达到的性能上限。

5.2.2 链接预测性能。 我们还评估了 RGLM 在有监督链接预测任务上的表现。遵循相关工作 [3, 32],，我们从节点分类训练集中随机选取节点对用于训练，并从测试集中选取节点对用于评估，确保边级训练集的规模与节点级训练集相同。如表2所示，RGLM 在链接预测任务中的表现始终优于所有基线方法。与仅

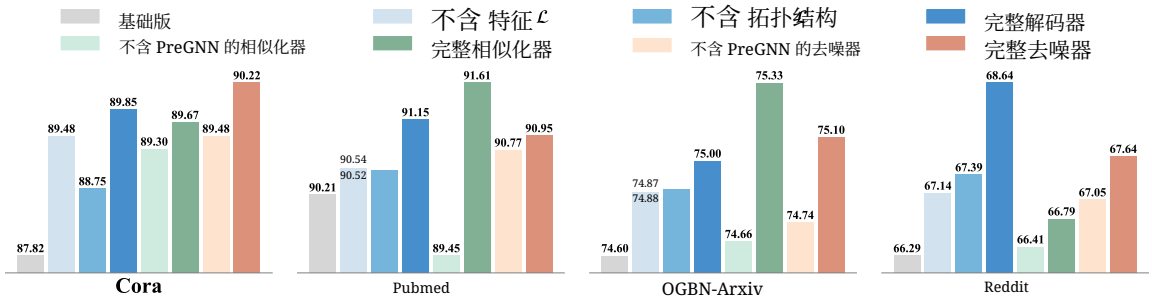


图4：不同数据集上的节点分类任务消融研究，其中报告了各变体的准确率。

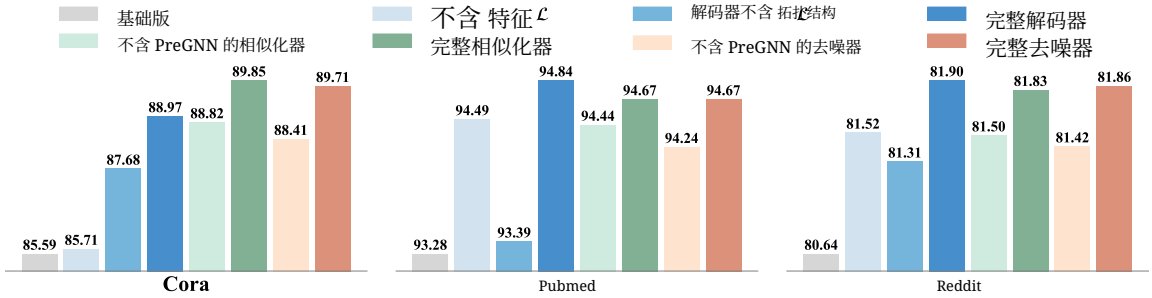


图5：不同数据集上的链接预测任务消融研究，各变体的结果以准确率报告。

表3：多数据集训练下的节点分类结果。最佳结果以粗体标出，次优结果加下划线。

方法	训练集	Cora+Pubmed			
	测试集	Cora		Pubmed	
	指标	Acc	F1	Acc	F1
基线方法	GCN	85.61	84.16	85.07	84.29
	GraphSAGE	86.35	86.21	89.27	89.12
	GAT	83.03	82.29	86.51	85.99
	UniMP	86.35	85.21	89.20	89.14
	NodeFormer	87.30	85.44	88.44	88.22
	GRACE	88.56	87.55	86.21	85.92
	GraphGPT	81.59	79.83	78.21	77.45
	LLaGA	88.53	87.28	89.60	89.46
Ours	RGLM-Decoder	89.67	88.14	90.37	90.32
	RGLM-Similarizer	88.93	87.49	**91.18**	**91.07**
	RGLM-Denoiser	**90.04**	**88.94**	90.34	90.36

利用文本监督的 GraphGPT 和 LLaGA 相比，RGLM 在对齐过程中明确引入了图监督，使准确率提升了 2.1 至 30 个百分点。这凸显了显式图监督对于促进图文对齐的关键作用。

5.2.3 多数据集泛化能力. 通常，在多个数据集上训练模型可能导致灾难性遗忘 [21]，从而使性能下降。在此，我们考察 RGLM 在多个数据集上的多数据集泛化能力。在我们的研究中，模型首先在 Cora 和 Pubmed 上针对节点分类任务进行训练，然后在每个数据集上进行评估。如表 3 所示，RGLM 在跨不同数据集训练时仍保持稳健，取得了更优的性能，并展现了其跨数据集学习能力。

5.2.4 零样本能力. 我们研究 RGLM 的零样本能力，即模型在源数据集上训练后，在训练期间无法接触目标数据集的情况下，直接在目标数据集上进行评估

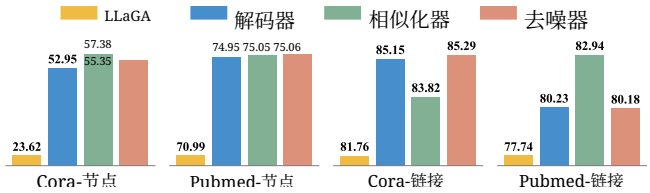


图 6：节点分类与链接预测任务上的零样本准确率，其中模型在 OGBN-Arxiv 上训练，并在未见过的 Cora 和 Pubmed 上评估。

。遵循 LLaGA [3]，，我们使用 OGBN-Arxiv 上的节点描述和链接预测任务训练模型，并分别使用 Cora 和 Pubmed 上的节点分类和链接预测任务对其进行评估。如图 6 所示，三种 RGLM变体在这两项任务中都展现出稳健的零样本能力。通过引入额外的图重构任务，知识可以在不同场景之间得到更有效的迁移。

5.3 模型分析

5.3.1 消融研究。我们首先通过消融研究评估各关键组件的重要性。图 4 和图 5 分别展示了三个 RGLM 变体在节点分类和链接预测任务上的消融结果。具体而言，“Vanilla”表示不使用任何重建损失的设置，大语言模型仅通过文本监督进行调优。“w/o L_{feat}”和“w/o L^{topo}”分别表示移除特征解码器和结构解码器。“w/oPreGNN”表示移除预训练图神经网络，这相当于在重构输入特征时忽略图上的结构信息。根据结果，我们观察到：（1）与纯文本监督相比，在指令微调过程中明确引入图监督能够持续提升性能，证明了 RGLM 的有效性。（2）对于解码器，同时引入特征重构和结构重构能够持续提升性能，且两者结合时效果最佳。

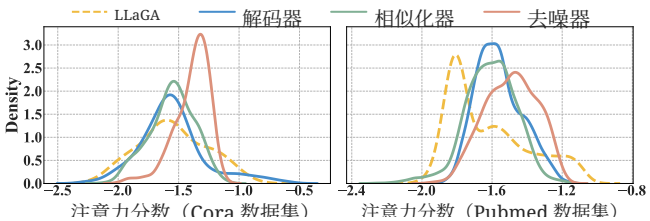


图 7：不同方法在图令牌上的注意力分布，其中注意力分数经过对数缩放。

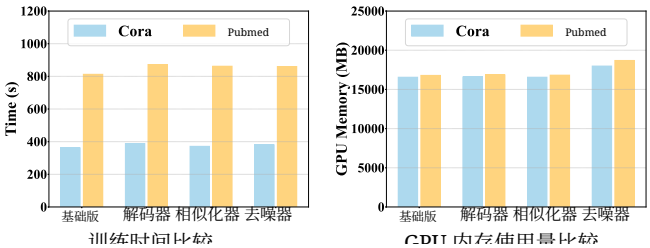


图 8：不同 RGLM 变体的时间复杂度（左）和资源开销（右）分析。

（3）对于相似化器和去噪器，即使不使用预训练图神经网络，RGLM 在大多数情况下仍能带来性能提升，这是因为它们近似退化为输入特征重建。

5.3.2 注意力分析。我们在 Cora 和 Pubmed 的测试集上，计算最后一个令牌相对于所有图令牌的注意力分数（不包括占位令牌）。图 7 展示了 LLaGA 和 RGLM 对图令牌的注意力分布。结果表明，RGLM 的注意力分数显著高于 LLaGA，这说明引入图重建目标能够有效引导大语言模型关注输入图，并增强其理解图的能力。

5.3.3 兼容性分析。我们分别采用 LLaMA3- 8B [33] 和 Vicuna-13B [6] 作为基础大语言模型，以评估 RGLM 的兼容性。我们报告了不同大语言模型在两种设置下的准确率：不使用重建损失（"基础版"）以及采用所提出的 RGLM。如图 9 所示，与纯文本监督相比，引入图重建损失能够持续提升不同大语言模型的性能，这表明 RGLM 与多种大语言模型兼容，并可作为通用流程。

5.3.4 复杂度分析。 我们进一步评估了 RGLM 的时间和空间复杂度。由于图重构损失不会影响大语言模型的推理阶段，因此这里我们重点关注指令微调阶段。图8报告了三种变体在 Cora 和 Pubmed 数据集上的训练时间与 GPU 内存开销。结果表明，重构过程仅带来了微不足道的训练时间和资源开销，而性能提升显然远超这点额外开销。

5.3.5 超参数分析。 最后，我们对重建损失系数进行敏感性分析：解码器的 λ_f 和 λ_s，以及相似化器和去噪器的 λ_l。如图10和

图11，RGLM的性能先上升后下降

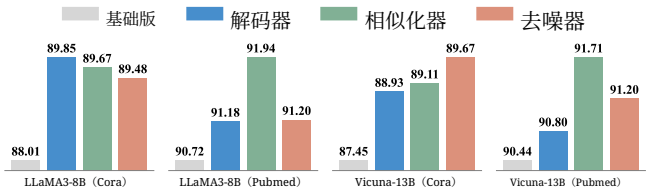


图9：不同 RGLM变体在基础大语言模型上的兼容性分析，其中报告了每种变体的准确率。

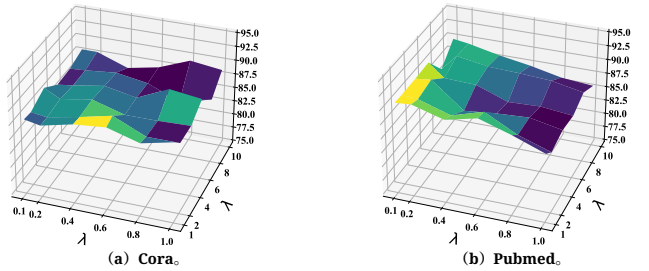


图10：RGLM-Decoder变体中超参数 λ_s 和 λ_f 的分析。

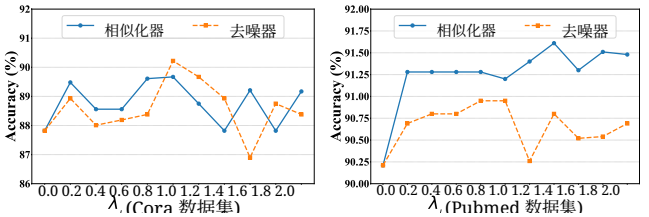


图11：RGLM-Similarizer 和 RGLM-Denoiser 变体中的超参数 λ_l 分析。

随着重建损失权重的增大。这一趋势源于重建目标作为正则化项的作用：当权重过小时，显式图监督变得不足，导致图信息利用不足；当权重过大时，重建损失主导训练并干扰语言建模，从而降低性能。

6 结论

在本研究中，我们从理论上揭示了现有 GTokenLLMs 的局限性：它们完全依赖语言指令提供的文本监督，从而只能实现隐式的图文对齐。为克服这一问题，我们证明，对齐目标的上界由输入图与其在大语言模型中的隐藏表示之间的互信息决定，并进一步提出重构式图指令微调流程 RGLM，以提高这一上界。RGLM 通过三种不同变体实现：在输入空间中使用解码器，在潜在空间中使用相似化器和去噪器。大量实验验证了 RGLM 的有效性。

致谢

本研究部分由国家自然科学基金委员会（编号：62550138、62192784、62572064、62472329、62322203、62172052）以及北京邮电大学优秀博士生基金（编号：CX20251005）资助。

参考文献

[1] Josh Achiam、Steven Adler、Sandhini Agarwal 等。2023。GPT-4 技术报告。arXiv 预印本 arXiv:2303.08774（2023）。[2] Jinwen Chen、Hao Miao、Dazhuo Qiu 等。2025。面向可持续性的任务推荐：空间众包中的应用。载于 ICDE。[3] Runjin Chen、Tong Zhao、Ajay Kumar Jaiswal 等。2024。LLaGA：大语言与图助手。载于国际机器学习会议。PMLR。[4] Xi Chen、Yan Duan、Rein Houthooft 等。2016。InfoGAN：通过信息最大化生成对抗网络实现可解释的表示学习。载于《神经信息处理系统进展》第29卷（2016）。[5] Zhikai Chen、Haitao Mao 等。2024。探索大语言模型（LLM）在图上学习中的潜力。载于《ACM SIGKDD探索通讯》（2024）。[6] Wei-Lin Chiang、Zhuohan Li、Ziqing Lin 等。2023。Vicuna：一款以媲美 GPT-4 的表现（达到 ChatGPT 质量的 90%*）令人瞩目的开源聊天机器人。见 https://vicuna. lmsys. org（访问于2023年4月14日）2、3（2023），6。[7] Jacob Devlin、Ming-Wei Chang 等。2019。BERT：用于语言理解的深度双向 Transformer 预训练。载于计算语言学协会北美分会。[8] Bahare Fatemi、Jonathan Halcrow 和 Bryan Perozzi。2024。像图一样说话：为大语言模型编码图。载于第十二届国际学习表征会议。[9] Matthias Fey 和 Jan Eric Lenssen。2019。使用 PyTorch Geometric 进行快速图表示学习。arXiv 预印本 arXiv:1903.02428（2019）。[10] Jiayan Guo、Lun Du、Hengyu Liu 等。2023。Gpt4graph：大语言模型能理解图结构数据吗？实证评估与基准测试。arXiv 预印本 arXiv:2305.15066（2023）。[11] Yuxin Guo、Deyu Bo、Cheng Yang、Zhiyuan Lu、Zhongjian Zhang、Jixi Liu、Yufei Peng 和 Chuan Shi。2024。以数据为中心的图学习：综述。《IEEE大数据汇刊》11卷1期（2024），1–20。[12] Zeyuan Guo、Enmao Diao、Cheng Yang 和 Chuan Shi。 [未注明年份。]用于连接图与 Transformer 的图词元化。载于第十四届国际学习表征会议。[13] Will Hamilton、Zhitao Ying 和 Jure Leskovec。2017。大规模图上的归纳式表示学习。载于《神经信息处理系统进展》（2017）。[14] Yufei He、Yuan Sui、Xiaoxin He 等。2025。UniGraph：面向文本属性图的统一跨领域基础模型学习。载于《第31届 ACM SIGKDD 知识发现与数据挖掘会议论文集》第1卷。[15] R Devon Hjelm、Alex Fedorov、Samuel Lavoie-Marchildon 等。2019。通过互信息估计与最大化学习深度表示。载于国际学习表征会议。[16] Edward J Hu、Yelong Shen、Phillip Wallis 等。2022。LoRA：大语言模型的低秩适配。ICLR 1卷2期（2022），3。[17] Weihua Hu、Matthias Fey、Marinka Zitnik 等。2020。开放图基准：用于图机器学习的数据集。载于《神经信息处理系统进展》（2020）。[18] Xuanwen Huang、Kaiqiao Han、Yang Yang 等。2024。图神经网络能否成为大语言模型的优秀适配器？载于《2024年 ACM万维网会议论文集》。[19] Bowen Jin、Gang Liu 等。2024。图上的大语言模型：全面综述。《IEEE知识与数据工程汇刊》（2024）。[20] Thomas N. Kipf 和 Max Welling。2017。基于图卷积网络的半监督分类。载于国际学习表征会议。[21] James Kirkpatrick、Razvan Pascanu、Neil Rabinowitz 等。2017。克服神经网络中的灾难性遗忘。载于《美国国家科学院院刊》（2017）。[22] Lecheng Kong、Jiarui Feng、Hao Liu 等。2025。GOFA：用于联合图语言建模的生成式通用模型。载于第十三届国际学习表征会议。[23] Guohao Li、Chenxin Xiong、Ali Thabet 等。2020。DeeperGCN：训练更深层图卷积网络所需的一切。arXiv 预印本 arXiv:2006.07739（2020）。[24] Yuhan Li、Peisong Wang、Xiao Zhu 等。2024。GLBench：面向大语言模型与图的综合基准。《神经信息处理系统进展》第37卷（2024），42349–42368。[25] Jiawei Liu、Cheng Yang、Zhiyuan Lu 等。2025。图基础模型：概念、机遇与挑战。《IEEE模式分析与机器智能汇刊》（2025）。[26] Yinghui Liu、Hao Miao、Guojiang Shen 等。2025。SPOT-Trip：双重偏好驱动的城外出行推荐。载于 NeurIPS。[27] Renqiang Luo、Huafei Huang、Ivan Lee 等。2025。FairGP：一种可扩展且公平的图Transformer，采用图划分技术。载于第39届 AAAI 人工智能会议论文集。[28] Renqiang Luo、Huafei Huang、Shuo Yu 等。2024。FairGT：公平性感知图Transformer。载于第32届国际人工智能联合会议论文集。[29] Ladislav Rampášek、Michael Galkin 等。2022。通用、强大且可扩展的图Transformer实践指南。载于《神经信息处理系统进展》（2022）。[30] Nils Reimers 和 Iryna Gurevych。2019年。Sentence-BERT：使用孪生 BERT 网络的句子嵌入。载于 2019年自然语言处理实证方法会议和第九届自然语言处理国际联合会议（EMNLP-IJCNLP）论文集。[31] Yunsheng Shi、Zhengjie Huang、Shikun Feng 等。2021年。掩码标签预测：用于半监督分类的统一消息传递模型。人工智能国际联合会议组织。[32] Jiabin Tang、Yuhao Yang、Wei Wei 等。2024年。Graphgpt：面向大语言模型的图指令微调。载于第47届国际 ACM SIGIR 信息检索研究与发展会议论文集。[33] Hugo Touvron、Thibaut Lavril、Gautier Izacard 等。2023年。Llama：开放且高效的基础语言模型。arXiv 预印本 arXiv:2302.13971（2023年）。[34] Ashish Vaswani、Noam Shazeer、Niki Parmar 等。2017年。注意力就是你所需的一切。神经信息处理系统进展（2017年）。[35] Petar Veličković、Guillem Cucurull、Arantxa Casanova 等。2018年。图注意力网络。载于国际学习表征会议。[36] Duo Wang、Yuan Zuo、Fengzhi Li 等。2024年。大语言模型作为零样本图学习器：图神经网络表征与大语言模型词元嵌入的对齐。神经信息处理系统进展 37（2024年），5950–5973。[37] Heng Wang、Shangbin Feng、Tianxing He 等。2023年。语言模型能用自然语言解决图问题吗？神经信息处理系统进展 36（2023年），30840–30861。[38] Haochen Wang、Yucheng Zhao、Tiancai Wang、Haoqiang Fan、Xiangyu Zhang 和 Zhaoxiang Zhang。2025年。Ross3d：具备三维感知的重建式视觉指令微调。载于 IEEE/CVF 国际计算机视觉会议论文集。9275–9286。[39] Haochen Wang、Anlin Zheng、Yucheng Zhao、Tiancai Wang、Zheng Ge、Xiangyu Zhang 和 Zhaoxiang Zhang。 [无日期。]。重建式视觉指令微调。载于第十三届国际学习表征会议。[40] Haoyu Peter Wang、Shikun Liu、Rongzhe Wei 等。2025年。大语言模型对文本属性图进行推理的泛化原理。载于第四十二届国际机器学习会议。[41] Jianing Wang、Junda Wu、Yupeng Hou 等。2024年。Instructgraph：通过以图为中心的指令微调和偏好对齐增强大语言模型。arXiv 预印本 arXiv:2402.08785（2024年）。[42] Fei Wu、Xiao-Yuan Jing、Pengfei Wei 等。2022年。应用于网页分类的半监督多视图图卷积网络。Information Sciences 591（2022年），142–154。[43] Qitian Wu、Chenxiao Yang、Wentao Zhao 等。2023年。DIFFormer：由能量约束扩散诱导的可扩展（图）Transformer。arXiv 预印本 arXiv:2301.09474（2023年）。[44] Qitian Wu、Wentao Zhao、Zenan Li 等。2022年。NodeFormer：用于节点分类的可扩展图结构学习变换器。神经信息处理系统进展（2022年）。[45] Zonghan Wu、Shirui Pan、Fengwen Chen 等。2020年。图神经网络综述。IEEE 神经网络与学习系统汇刊（2020年）。[46] Bo Yan、Yang Cao、Haoyu Wang 等。2024年。用于隐私保护推荐的联邦异构图神经网络。载于2024年 ACM万维网会议论文集。[47] Bo Yan、Sihao He、Cheng Yang 等。2025年。基于信息瓶颈原理的联邦图浓缩。载于 AAAI 人工智能会议论文集。[48] Bo Yan、Zhongjian Zhang、Huabin Sun、Mengmei Zhang、Yang Cao 和 Chuan Shi。2025年。基于大语言模型的以数据为中心的联邦图学习。arXiv 预印本 arXiv:2503.19455（2025年）。[49] Jinyu Yang、Ruijia Wang 等。2025年。利用语言模型实现跨异质性图知识迁移。载于 AAAI 人工智能会议论文集。[50] Ling Yang、Ye Tian、Minkai Xu 等。2024年。VQGraph：重新思考图表征空间，以连接图神经网络与多层感知机。载于第十二届国际学习表征会议。[51] Zhilin Yang、William Cohen 和 Ruslan Salakhudinov。2016年。重新审视基于图嵌入的半监督学习。载于国际机器学习会议。[52] Ruosong Ye、Caiqi Zhang、Runhui Wang 等。2024年。语言是图所需的一切。载于 EACL（论文集）。[53] Yuning You、Tianlong Chen、Yongduo Sui 等。2020年。结合增强的图对比学习。神经信息处理系统进展（2020年）。[54] Chaohao Yuan、Kangfei Zhao、Ercan Engin Kuruoglu 等。2025年。图变换器综述：架构、理论与应用。arXiv 预印本 arXiv:2502.16533（2025年）。[55] Mengmei Zhang、Mingwei Sun、Peng Wang 等。2024年。Graphtranslator：将图模型与大语言模型对齐，以处理开放式任务。载于2024年 ACM万维网会议论文集。1003–1014。[56] Zhongjian Zhang、Xiao Wang 等。2025年。大语言模型能否提升图神经网络的对抗鲁棒性？载于第31届 ACM SIGKDD 知识发现与数据挖掘会议论文集，第1卷。

[57] Zhongjian Zhang、Mengmei Zhang 等。2024年。赋予预训练图模型可证明的公平性。载于 *2024*年 *ACM*万维网会议论文集。[58] Zhongjian Zhang、Mengmei Zhang、Xiao Wang 等。2025年。从稀疏聚合视角重新思考联邦推荐中的拜占庭鲁棒性。载于 *AAAI* 人工智能会议论文集。[59] Jianan Zhao、Le Zhuo、Yikang Shen 等。2024年。GraphText：文本空间中的图推理。载于 自适应基础模型：面向个性化与高效学习的演进式人工智能。

[60] Cai Zhou、Xiyuan Wang 和 Muhan Zhang。2024年。通过潜在图扩散统一图上的生成与预测。神经信息处理系统进展 (2024年)。[61] Huachi Zhou、Jiahe Du 等。2025年。Eachgraph 是一种新语言：使用大语言模型进行图学习。载于计算语言学协会：*ACL 2025*年会议论文集。计算语言学协会。[62] Yanqiao Zhu、Yichen Xu 等。2020年。深度图对比表示学习。载于*ICML* 图表示学习及其他主题研讨会。

A 相关工作

A.1 图上的学习

图数据在现实世界中无处不在，包括社交网络 [18, 57], 网页网络 [42], 引文网络 [47, 51], 和电子商务网络 [2, 26, 46, 58]。然而，由于图本身具有不规则性和复杂的连接模式，有效建模图结构是一项根本性挑战。图神经网络（GNNs）已成为应对这一问题的强大工具，它们利用消息传递机制从局部邻域聚合信息 [11, 45]。GCN [20], 、GAT [35] 和 GraphSAGE [13] 等代表性模型在节点分类和链接预测任务中取得了显著成功。与此同时，将基于 Transformer 的架构 [34]应用于图领域 [27, 28, 43, 44, 54]也日益受到关注。这些模型旨在捕获全局依赖和更丰富的结构语义，为传统局部聚合方案提供互补的视角。尽管取得了这些进展，现有图神经网络的一项关键局限在于任务泛化能力较弱。大多数现有模型以特定任务和数据集为依托进行训练，这阻碍了它们在迁移到不同分布或任务时的性能。

A.2 面向图的大语言模型

近期，大语言模型的卓越成功 [1,6, 33] 促使研究人员将其应用于图学习领域，最终目标是开发能够在多种场景中泛化的图基础模型 [12, 25, 48, 49]。通常，大语言模型被用作预测器来完成各种图任务，因此受到广泛关注 [19]。一项关键挑战在于如何将图数据与自然语言嵌入空间对齐，使大语言模型能够更好地理解图。为应对这一挑战，现有研究主要分为图文本化大语言模型（GTextLLMs）和图标记化大语言模型（GTokenLLMs）。

GTextLLMs [5, 8, 10, 37, 41, 59] 利用自然语言通过人工设计的提示模板描述图，并直接与文本对齐。例如，NLGraph [37] 将图表示为嵌入文本模板中的节点与边列表。GraphText [59]使用句法图解析，将图转换为基于树的文本格式。然而，这类策略往往存在冗余和表述不准确的问题 [3, 5, 32]。相比之下，GTokenLLMs [3, 22, 32, 36, 55] 将复杂结构和长文本编码为图标记序列，并将其作为查询大语言模型的前缀令牌。为了将图令牌与文本嵌入空间对齐，它们通常会针对图任务构建各种语言指令，并学习一个投影器以促进模态对齐。例如，GraphGPT [32] 围绕图匹配任务构建语言指令，而 LLaGA [3] 则通过节点分类、链接预测和节点描述任务构建指令。尽管取得了这些成功，它们的对齐过程仍采用以大语言模型为中心的设计，仅依赖文本监督，导致图文对齐仅为隐式对齐，从而限制了性能。

B 证明

B.1 命题 4.1 的证明

证明。由于 $\mathbf{H}$ 是 $\boldsymbol{s}^G$ 的确定性函数（即均值池化），数据处理不等式保证 $\mathrm{I}(\mathrm{G};\boldsymbol{s}^G) \geq \mathrm{I}(\mathrm{G};\mathbf{H})$。通过将 $\mathbf{H}$ 解码回节点特征 $\mathbf{Z}$ 和邻接矩阵 $\mathbf{A}$，其中 $\mathbf{A}_{ij}=1$ 表示节点 v_i 和 v_j 相连，否则 $\mathbf{A}_{ij}=0$，重建损失可表示为：

$$\mathcal{L}_{\text{graph}} = \lambda_f \,\mathbb{E}\big[-\log q_f(\mathbf{Z}|\mathbf{H})\big] + \lambda_s \,\mathbb{E}\big[-\log q_s(\mathbf{A}|\mathbf{H})\big]. \tag{29}$$

假设 $q_f(\mathbf{Z}|\mathbf{H}) = \mathrm{N}(\mu_f(\mathbf{H}), \sigma^2\mathbf{I})$，这对应于特征上的均方误差损失；以及 $q_s(\mathbf{A}|\mathbf{H}) = \prod_{i<j}\mathrm{Bernoulli}(\pi_s(\mathbf{h}_i,\mathbf{h}_j))$，这对应于边上的二元交叉熵损失。根据 Barber–Agakov 恒等式，对于任意条件模型 q，我们有：

$$\begin{aligned}\mathbb{E}[-\log q_f(\mathbf{Z}|\mathbf{H})] &= \mathcal{H}(\mathbf{Z}|\mathbf{H}) + \mathbb{E}[\mathrm{KL}(p(\mathbf{Z}|\cdot)\,\|\,q_f(\mathbf{Z}|\cdot))] \\ &\geq \mathcal{H}(\mathbf{Z}|\mathbf{H}), \end{aligned}\tag{30}$$

$$\begin{aligned}\mathbb{E}[-\log q_s(\mathbf{A}|\mathbf{H})] &= \mathcal{H}(\mathbf{A}|\mathbf{H}) + \mathbb{E}[\mathrm{KL}(p(\mathbf{A}|\cdot)\,\|\,q_s(\mathbf{A}|\cdot))] \\ &\geq \mathcal{H}(\mathbf{A}|\mathbf{H}). \end{aligned}\tag{31}$$

因此，

$$\mathcal{L}_{\text{graph}} \;\geq\; \lambda_f\mathcal{H}(\mathbf{Z}|\mathbf{H}) + \lambda_s\mathcal{H}(\mathbf{A}|\mathbf{H}). \tag{32}$$

在此，我们假设图是一对独立的随机变量 $(\mathbf{Z},\mathbf{A})$。互信息可以分解为：

$$\begin{aligned}\mathcal{I}(\mathcal{G};\mathbf{H}) &= \mathcal{I}(\mathbf{Z};\mathbf{H}) + \mathcal{I}(\mathbf{A};\mathbf{H}) \\ &= \mathcal{H}(\mathbf{Z}) - \mathcal{H}(\mathbf{Z}|\mathbf{H}) + \mathcal{H}(\mathbf{A}) - \mathcal{H}(\mathbf{A}|\mathbf{H}). \end{aligned}\tag{33}$$

由于 $\mathrm{H}(\mathbf{Z})$ 和 $\mathrm{H}(\mathbf{A})$ 与参数无关，结合上述不等式并舍弃非负的 KL 项，可得：

$$\mathcal{I}(\mathcal{G};\boldsymbol{s}^G) \;\geq\; \mathcal{H}(\mathbf{Z}) + \mathcal{H}(\mathbf{A}) \;-\; \tfrac{1}{\lambda_f}\mathcal{L}_{\text{feat}} \;-\; \tfrac{1}{\lambda_s}\mathcal{L}_{\text{topo}}. \tag{34}$$

□

该证明表明，最小化原始图重建损失可最大化 $\mathrm{I}(\mathrm{G};\boldsymbol{s}^G)$ 的一个可计算下界，从而直接改善优化目标 C-GTMI $\mathrm{I}(\boldsymbol{x}_t;\boldsymbol{s}^G|\boldsymbol{s}^T_{<t})$ 的上界。

B.2 命题 4.2 的证明

证明。给定一个子图 G，令 $\mathbf{E} = g(\mathrm{G})$ 表示由预训练图神经网络 g提取的潜在节点表示矩阵。重建令牌$\mathbf{H}$通过轻量级多层感知机 $s(\cdot)$ 映射到 $\mathbf{E}$，并通过归一化确保 $\|\mathbf{E}\| = \|s(\mathbf{H})\| = 1$。重建损失定义为余弦相似度：

$$\mathcal{L}_{\text{sim}} = \mathbb{E}[1 - \cos(s(\mathbf{H}), \mathbf{E})]. \tag{35}$$

在单位球面上，我们假设$\mathbf{E}$ 在给定$\mathbf{H}$的条件下服从 von Mises–Fisher（vMF）分布：

$$q_s(\mathbf{E}|\mathbf{H}) = \mathrm{vMF}(\mathbf{E};\,\mu(\mathbf{H}),\kappa) \propto \exp\big(\kappa\cdot\lambda_l\cdot\cos(\mu(\mathbf{H}),\mathbf{E})\big), \tag{36}$$

其中 $\mu(\mathbf{H}) := \frac{s(\mathbf{H})}{\|s(\mathbf{H})\|}$, $\kappa>0$, λ_l是一个缩放因子。因此：

$$-\log q_s(\mathbf{E}|\mathbf{H}) = -\kappa\cdot\lambda_l\cdot\cos(s(\mathbf{H}),\mathbf{E}) + \log C_d(\kappa), \tag{37}$$

其中，$\log C_d(\kappa)$ 是一个与 $\mathbf{H}$ 无关的常数。因此，

$$\begin{aligned}\mathbb{E}[-\log q_s(\mathbf{E}|\mathbf{H})] &= \text{const} + \kappa\cdot\lambda_l\cdot\mathbb{E}[1-\cos(s(\mathbf{H}),\mathbf{E})] \\ &= \text{const} + \kappa\cdot\lambda_l\cdot\mathcal{L}_{\text{sim}}. \end{aligned}\tag{38}$$

对于任意条件模型 $q_s(\mathbf{E}|\mathbf{H})$，我们有：

$$\begin{aligned}\mathbb{E}[-\log q_s(\mathbf{E}|\mathbf{H})] &= \mathcal{H}(\mathbf{E}|\mathbf{H}) + \mathbb{E}[\mathrm{KL}(p(\mathbf{E}|\cdot)\,\|\,q_s(\mathbf{E}|\cdot))] \\ &\geq \mathcal{H}(\mathbf{E}|\mathbf{H}). \end{aligned}\tag{39}$$

因此，我们得到：

$$\mathcal{L}_{\text{sim}} \geq \frac{1}{\kappa \cdot \lambda_l}\mathcal{H}(\mathbf{E}|\mathbf{H}) - \text{const} \Rightarrow \mathcal{H}(\mathbf{E}|\mathbf{H}) \leq \kappa \cdot \lambda_l \cdot \mathcal{L}_{\text{sim}} + \text{const}. \tag{40}$$

Since $\mathcal{I}(\mathbf{E};\mathbf{H}) = \mathcal{H}(\mathbf{E}) - \mathcal{H}(\mathbf{E}|\mathbf{H})$, we obtain:

$$\mathcal{I}(\mathbf{E};\mathbf{H}) \;\geq\; \mathcal{H}(\mathbf{E}) - (\kappa \cdot \lambda_l \cdot \mathcal{L}_{\text{sim}} + \text{const}). \tag{41}$$

此外，g 被冻结，因此 $\mathbf{E} = g(\mathbf{G})$ 是 G 的确定性函数，而 H 是 s^G 的确定性函数，因此我们有：

$$\mathcal{I}(\mathcal{G}; s^G) \geq \mathcal{I}(\mathcal{G};\mathbf{H}) \geq \mathcal{I}(\mathbf{E};\mathbf{H}) \geq \mathcal{H}(\mathbf{E}) - (\kappa \cdot \lambda_l \cdot \mathcal{L}_{\text{sim}} + \text{const}). \tag{42}$$

□

证明表明，最小化H与E之间的相似性损失，可以最大化 I(G; s^G) 的一个可计算下界，进而提升 C-GTMI I(x_t s^G | ;) 可达到的上界。 $s^T_{<t}$

B.3 命题 4.3 的证明

证明：

给定一个子图 G，令$\mathbf{E} = g(\mathbf{G})$表示由预训练图神经网络 g 提取的潜在节点表示矩阵。重建令牌$\mathbf{H}$ 被投影到条件矩阵$\mathbf{C} = \Psi(\mathbf{H})$中。目标是对条件分布$q_f(\mathbf{E}_{t-1}|\mathbf{E}_t, \mathbf{C})$进行建模，使得干净嵌入$\mathbf{E}$能够在条件$\mathbf{C}$下从其噪声对应项$\mathbf{E}_t$中恢复。前向扩散过程被表述为一个马尔可夫链：通过逐步添加高斯噪声来扰动干净嵌入$\mathbf{E}$，从而构造含噪嵌入$\mathbf{E}_t$：

$$\begin{aligned} p(\mathbf{E}_t|\mathbf{E}) &= \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\,\mathbf{E},\ (1-\bar{\alpha}_t)\,\mathbf{I}\big), \\ \mathbf{E}_t &= \sqrt{\bar{\alpha}_t}\mathbf{E} + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I}), \end{aligned} \tag{43}$$

其中，$\mathrm{N}(\boldsymbol{\mu}, \Sigma)$表示高斯分布，$t$表示离散时间步，$\alpha_t = 1 - \beta_t$、$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_t$和 $\beta_t \in (0,1)$是预定义的方差调度。潜在扩散模型学习反转这一逐步添加噪声的过程。具体而言，为了在以$\mathbf{C}$为条件的情况下，从纯噪声$\mathbf{E}_t$迭代生成干净嵌入$\mathbf{E}$，反向过程被参数化为

$$\begin{aligned} q_f(\mathbf{E}_{t-1}|\mathbf{E}_t, \mathbf{C}) &= \mathcal{N}\Big(\boldsymbol{\mu}_f(\mathbf{E}_t, \mathbf{C}, t),\ \sigma_t^2\mathbf{I}\Big), \\ \boldsymbol{\mu}_f &= \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{E}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\boldsymbol{\epsilon}_f(\mathbf{E}_t; \mathbf{C}, t)\right). \end{aligned} \tag{44}$$

边缘条件似然由下式给出

$$\begin{aligned} \log q_f(\mathbf{E}|\mathbf{H}) = \log \int q_f(\mathbf{E}_T) \prod_{t=1}^{T} q_f(\mathbf{E}_{t-1}|\mathbf{E}_t, \mathbf{C}) \\ \times\, q_f(\mathbf{E}|\mathbf{E}_1, \mathbf{C})\, d\mathbf{E}_{1:T}. \end{aligned} \tag{45}$$

通过引入前向分布 $p(\mathbf{E}_{1:T}|\mathbf{E})$ 作为变分近似，可以推导出标准变分下界（证据下界）：

$$\begin{aligned} \log q_f(\mathbf{E}|\mathbf{H}) \geq\ & -\mathrm{KL}\left(p(\mathbf{E}_T|\mathbf{E}) \,\|\, q_f(\mathbf{E}_T)\right) \\ & - \sum_{t=2}^{T} \mathbb{E}\left[\mathrm{KL}\big(p(\mathbf{E}_{t-1}|\mathbf{E}_t, \mathbf{E}) \,\|\, q_f(\mathbf{E}_{t-1}|\mathbf{E}_t, \mathbf{C})\big)\right] \\ & + \mathbb{E}\left[\log q_f(\mathbf{E}|\mathbf{E}_1, \mathbf{C})\right]. \end{aligned} \tag{46}$$

在方差保持调度和高斯参数化下，每个 KL项都可以改写为均方误差（MSE）形式的噪声预测，其差异仅在于一个与参数无关的常数：

$$\begin{aligned} \mathrm{KL}\big(p(\mathbf{E}_{t-1}|\mathbf{E}_t, \mathbf{E}) \| q_f(\mathbf{E}_{t-1}|\mathbf{E}_t, \mathbf{C})\big) = \tilde{\lambda}_t \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \|\boldsymbol{\epsilon}_f(\mathbf{E}_t; \mathbf{C}, t) - \boldsymbol{\epsilon}\|_2^2 \\ + \text{const}_t, \end{aligned} \tag{47}$$

其中 $\tilde{\lambda}_t > 0$ 取决于噪声调度。将其代回证据下界并合并常数项，可得扩散损失，它与负证据下界仅相差一个常数：

$$\mathcal{L}_{\text{diff}} = \sum_{t=1}^{T} \lambda_t\, \mathbb{E}_{(\mathbf{E},\mathbf{H}),\,\boldsymbol{\epsilon}} \Big\| \boldsymbol{\epsilon}_f\big(\sqrt{\bar{\alpha}_t}\mathbf{E} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon};\ \mathbf{C}, t\big) - \boldsymbol{\epsilon} \Big\|_2^2, \tag{48}$$

其中 $\lambda_t > 0$ 将常数吸收到权重中。这种等价关系给出了负对数似然的上界：

$$\mathbb{E}[-\log q_f(\mathbf{E}|\mathbf{H})] \;\leq\; \text{const} + \lambda_l \cdot \mathcal{L}_{\text{diff}}, \tag{49}$$

其中 λ_l 是一个缩放因子。此外，对于任意条件模型$q_f(\mathbf{E}|\mathbf{H})$，有：

$$\begin{aligned} \mathbb{E}[-\log q_f(\mathbf{E}|\mathbf{H})] &= \mathcal{H}(\mathbf{E}|\mathbf{H}) + \mathbb{E}\big[\mathrm{KL}\big(p(\mathbf{E}|\mathbf{H}) \,\|\, q_f(\mathbf{E}|\mathbf{H})\big)\big] \\ &\geq \mathcal{H}(\mathbf{E}|\mathbf{H}). \end{aligned} \tag{50}$$

舍去非负的KL项，可得条件熵的如下上界：

$$\mathcal{H}(\mathbf{E}|\mathbf{H}) \;\leq\; \text{const} + \lambda_l \cdot \mathcal{L}_{\text{diff}}. \tag{51}$$

根据 $\mathrm{I}(\mathbf{E};\mathbf{H}) = \mathrm{H}(\mathbf{E}) - \mathrm{H}(\mathbf{E}|\mathbf{H})$ 的定义，我们得到：

$$\mathcal{I}(\mathbf{E};\mathbf{H}) \;\geq\; \mathcal{H}(\mathbf{E}) - (\text{const} + \lambda_l \cdot \mathcal{L}_{\text{diff}}). \tag{52}$$

最后，由于 $\mathbf{E} = g(\mathbf{G})$ 是 G 的确定性函数，且 H 是通过对 s^G 进行均值池化确定性地得到的，因此我们有：

$$\mathcal{I}(\mathcal{G}; s^G) \geq \mathcal{I}(\mathbf{E}; s^G) \geq \mathcal{I}(\mathbf{E};\mathbf{H}) \geq \mathcal{H}(\mathbf{E}) - (\text{const} + \lambda_l \cdot \mathcal{L}_{\text{diff}}). \tag{53}$$

□

证明表明，最小化扩散噪声预测损失 L_{diff}可最大化 I(G; s^G) 的可计算下界，进而提升 C-GTMI $\mathrm{I}(\boldsymbol{x}_t;\ s^G \,|\, s^T_{<t})$ 可达到的上界。

C 实验细节

C.1 数据集描述

我们在四个广泛使用的数据集上开展实验：来自引文网络的 Cora [51],Pubmed [51] 和 OGBN-Arxiv [17]，以及来自社交网络的 Reddit [18]。数据集统计信息见表 4，详细描述如下：

• **Cora** 是一个引文网络，其中每个节点代表一篇计算机科学领域的研究论文，相关的原始文本特征包括论文的标题和摘要。边表示论文之间的引文关系。每个节点都被赋予一个标签，用于标明论文的类别。

• **Pubmed** 是一个由生物医学领域研究论文构成的引文网络，其中每个节点对应一篇论文，每条边表示两篇论文之间的引文关系。

• **OGBN-Arxiv** 是一个包含研究论文及其引文关系的引文网络，数据收集自 arXiv 存储库。每个节点代表一篇论文，每条边表示两篇论文之间的引文链接。

• **Reddit**是一个社交网络，其中每个节点代表一位用户，

表 4：实验数据集的统计信息。

数据集	# 节点	# 边	# 类别	划分	领域
Cora	2,708	5,429	7	6:2:2	引文
Pubmed	19,717	44,338	3	6:2:2	引文
OGBN-Arxiv	169,343	1,166,243	40	6:2:3	引文
Reddit	33,434	198,448	2	1:1:8	社交网络

边表示两位用户是否曾通过互相回复进行互动。与每个节点关联的原始文本由该用户此前发布的子版块帖子内容组成，最多包含最近的三条。每个节点都带有标签，标明该用户属于受欢迎还是普通类别。

C.2 基线描述

传统的图神经网络模型包括：

• **图卷积网络** [20] 是一种基于谱理论的受欢迎的图卷积网络。

• **图注意力网络** [35] 由多个注意力层组成，可以为不同的邻居节点学习不同的权重。

• **GraphSAGE** [13] 通过利用自我节点特征以及固定数量的邻近节点特征，归纳式地学习节点嵌入。

基于 Transformer 的图模型包括：

• **UniMP** [31] 采用图 Transformer 网络，在训练和推理过程中整合特征与标签传播。

• **NodeFormer** [44]使用固定阈值过滤不相似的邻居，并聚合多视图邻域信息以解决图不一致性。

• **DIFFormer** [43] 提出了一种受能量约束的扩散式图 Transformer，旨在通过融合其他实例的上下文信息，将一批实例编码为演化状态。

预训练-微调图神经网络包括：

• **GraphCL** [53] 采用多种图增强技术生成同一图的多个视图，并通过最大化这些视图之间的特征一致性来学习节点表示。

• **GRACE** [62] 提出通过随机边丢弃和随机节点特征丢弃来增强图结构，并使用受损图作为对比视图。

• **LGD** [60] LGD 先预训练一个自编码器，将图嵌入强大的潜在空间，再在该空间中训练扩散模型，以统一图生成和预测任务。

GTextLLMs 包括：

• **GraphText-ICL** [59] 是 GraphText 的上下文学习变体，无需调优过程。GraphText 将节点属性和关系封装在图语法树中，并由大语言模型对其进行处理，从而解耦深度与范围。

GTokenLLMs 包括：

• **GraphText-SFT** [59] 是 Graph-Text 的监督微调版本，它采用多层感知机投影器将节点连续特征映射到文本输入空间，以便查询大语言模型。

• **GraphGPT** [32] 使用预训练图神经网络对图数据进行编码，并引入双阶段指令调优，使图表示与自然语言对齐。

• **LLaGA** [3] 设计了两种图描述模板，用于序列化图数据，并通过三种不同的任务对大语言模型进行指令调优，包括节点描述、节点分类和链接预测任务。

C.3 实现细节

首先，RGLM 旨在将图监督显式融入调优过程，而不是设计新的指令数据集。因此，我们使用 LLaGA [3] 的指令数据集来训练模型。在调优阶段，我们使用 AdamW优化器训练模型一个 epoch，将每台设备的批大小设为 4，并将投影器学习率设为 2e-3。根据 LLaGA [3],的做法，为弥补数据量有限的问题，我们将最小数据集 Cora 和 Reddit 中的训练样本重复三次。预热比例设为 3e-2，大语言模型的最大输入长度为 4096。对于邻居细节模板，我们对每个节点采样两跳邻居，每跳的采样大小设为 10。我们采用 LoRA [16] 进行高效微调，学习率、丢弃率、排名和 α 分别设为 5e-4、0.1、8 和 32。在推理阶段，我们将大语言模型的温度设为 0.001，以确保结果具有可复现性。我们将所有三种 RGLM变体的学习率统一设为 5e-4，并按如下方式配置超参数：对于 RGLM-Decoder，超参数 λ_f从 {0.1, 0.2, 0.4,0.6, 0.8, 1.0} 中选取，λ_s从 {1,2, 4,6, 8, 10} 中选取。对于 RGLM-Similarizer 和 RGLM-Denoiser，λ_l从 0.2 到 2.0 以 0.2 为步长选取。

对于基于语言模型的预训练GNN编码器，按照 [29, 60],，我们采用 DeeperGCN [23] 和图注意力网络 [35] 的组合作为骨干网络，包含 7 层，隐藏维度为 96。我们使用拉普拉斯位置编码和随机游走结构编码。模型训练 500 轮，掩码标签比例为 0.8，并采用余弦学习率调度，预热 50 轮，初始学习率为 1e-3。

为便于公平比较，我们采用统一文本编码器 Sentence-BERT [30],，从原始文本属性中提取节点特征。所有 GTokenLLMs 均使用 Vicuna-7B-v1.5-16K [6] 作为基础大语言模型。对于这些基线模型，我们根据作者提供的源代码进行实现。除非另有说明，我们采用作者实现中的默认参数设置。对于未包含在已发布基线代码中的数据集，我们使用网格搜索策略调整所有基线模型。对于所有实验，我们在验证集上选择最优超参数，并将其应用于测试集。所有模型均使用 PyG [9]实现。

C.4 计算环境与资源 rces

实验在具有以下规格的计算环境中进行：

• **操作系统**：Linux Ubuntu 5.15.0-102-generic。• **CPU**: Intel(R) Xeon(R) Platinum 8358 CPU @ 2.60GHz。• **GPU**: NVIDIA A800 80GB。